To work or not to work:

The effect of response requirement variation on signal detection performance in hens

A thesis

submitted in partial fulfilment

of the requirements for the degree

of

Master of Applied Psychology (Behaviour Analysis)

at

The University of Waikato

by

ANNA KALIOPE TASHKOFF

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

2017

Abstract

The role of effort in an SDT paradigm has not been adequately investigated using only natural contingencies, where hits are the only reinforced responses. Fixed-ratio (FR) requirement as a measure of effort was systematically varied in a go/no-go signal detection task. Hens were trained to discriminate between a brighter keylight (S+) and a dimmer keylight (S-), where a fixed-ratio response requirement was in effect on S+ trials (i.e., for a "go" response) and a secondary, 'advance' key progressed to the next trial at any point following an observation response (i.e., a "no-go" response). A negative, linear relationship was discovered between FR requirement and hit rate. Although FR requirement variation was not found to significantly influence specificity performance, a graphical trend was observed such that, as FR increased, specificity generally increased before levelling off at a FR 16 response requirement. Comparisons between original and reinstated conditions suggest that performance was not affected by an order or practice effect. Implications and limitations of these findings are discussed, and considerations for future research are identified, such as generalisation of these findings across species and differing types of 'effort'.

Acknowledgements

To Professors Mary Foster and James McEwan, I give the greatest thanks for instilling in me a love for this discipline and providing countless hours guiding fledgling psychologists on our own personal journeys with enthusiasm, patience, and care. Truly great teachers do not come around often, and I will miss the opportunity to work with you. Ngā mihi whakawhetai.

This work was incomparably affected by the supervision of Dr. Timothy Edwards, who provided seemingly endless insight, surgical feedback, and positive support, always helping me rise to the challenge. Thank you for showing me the potential joy in research. Ngā mihi nunui.

Special thanks to Associate Professor John Perrone for helping make stats less scary! I appreciated all of your excellent advice. Ngā mihi.

The technical (and oftentimes emotional) assistance provided by Jenny Chandler and Rob Bakker was invaluable - I could not have done this without you, whether it was helping to fix experiment malfunctions or keeping the lab environment cheerful by blasting the radio or humming a ditty. Thanks for letting me bug you every time something went wrong! Kia ora rawa atu.

To all of the other people from the Animal Behaviour and Welfare Research Centre, staff and students alike, thank you for your contributions, advice, and encouragement; special thanks here to Karen Sluter for all of your gentle nudges in the right direction. Noho ora mai ki āu koutou ake haerenga.

To my family, thank you for always believing in me and banding around me when I struggled; it means more than you know. To my friends, you all mean the world to me, and I couldn't have done this without you. Lastly, to Tadhg, your love and support and late-night lab companionship are really the only reasons this thing managed to finally get finished. He aroha mutunga kore, mō ake tonu atu.

Table of Contents

List of Tables

List of Figures

The ability to correctly identify and discriminate between two or more stimuli is fundamental to many of the processes and challenges in life, for both humans and animals. As such, the factors involved in the process of stimulus discrimination and control have been of central importance in the field of behavioural psychology, examining differential responding in the presence or absence of certain stimuli (Shahan & Chase, 2002). In typical tasks of stimulus discrimination with operant behaviour, a subject must correctly perceive differences between one stimulus (i.e., the 'target') from one or more other, non-target stimuli. Subsequently, performing the correct behaviour in the presence of the target stimulus (S+) will lead to reinforcement, while doing so in the presence of the other stimuli (S-) will not (Jenkins, 1965; Shahan & Chase, 2002).

These stimuli may differ on a number of dimensions (e.g., colour, shape, size, luminosity, etc.) so as to render the difference between the S+ and S- discriminable, and these differences may vary the difficulty of a discrimination task. That is, the 'signal strength' of an S+ relative to an S- can vary between being high and easily discriminable (i.e., the differences between the S+ and S- are to a larger degree and/or across more dimensions) to being so low as to be indistinguishable, and discrimination would be reduced to guessing behaviour (i.e., zero signal strength; Nevin, 1969).

In initially describing the discriminative process, threshold theories held that there was a minimum level of signal strength intensity (i.e., an objective threshold) that a stimulus must surpass in order to be consciously recognised (Rouder & Morey, 2009). These discrete-state theories of stimulus discrimination held the existence of two, mutually exclusive mental states (i.e., detection versus non-detection); theoretically, stimulus presentation would elicit one of these states and lead to a related response (Malmberg, 2002; Rouder & Morey, 2009).

However, a failure to adequately present unifying principles using true thresholds while accounting for individual differences in decision making, perceptual ability, sensory state, and response processes led to controversy, and these theories fell somewhat out of favour in perceptual psychology (Campbell, 1964; Nevin, 1969). The development of signal detection theory meant that some of these issues were able to be addressed using a behavioural paradigm.

Signal Detection Theory

Signal detection theory (SDT) provides a framework for examining discrimination behaviour and decision-making behaviour, where the presence of the target signal or stimulus (S+) must be detected and differentiated from the background 'noise' (i.e., other environmental and internal stimuli, or S-; Abdi, 2007). The simplest discrimination task utilising SDT is called the yes-no task, where on each trial one stimulus condition (i.e., either with the S+, or the S- alone) is presented, and the subject must respond either 'yes' or 'no' to indicate the presence or absence of the S+, respectively (Stanislaw & Todorov, 1999). In SDT, the subject is theorised to respond either 'yes' or 'no' depending on an individually set criterion. A decision variable influences responses, such that when the salient characteristics of the stimulus (e.g., brightness) are perceived to exceed this criterion, the response would be 'yes,' with a 'no' response occurring if the criterion was not exceeded (Macmillan, 2002).

This decision variable reflects the inferred internal states that are occasioned upon stimulus presentation (Boneau & Cole, 1967). While stimulus presentation is a fixed, objective event, the physio-/psychological sensory mechanisms (i.e. private events) that occur during the processing of stimuli presentations are assumed to continuously vary, adding internal 'noise' to the

discrimination process; this is the fundamental difference between discrete-state theories and the continuous-state model of SDT (Malmberg, 2002).

The two stimulus conditions and two response alternatives result in a categorical matrix of performance measures based on response correctness, as shown in Table 1.

Table 1

SDT Categories of Responses

		Stimulus Condition	
		S- (S+ Absent)	S+ Present
Response	Yes	False Alarm	Hit
Alternative	No	Correct Rejection	Miss

As detailed in Table 1, responding 'yes' when the S+ is present is nominally categorised as a 'hit', and a 'no' response when the S+ is absent is labelled a 'correct rejection'; both of these responses are correct. For the incorrect response alternatives, a 'false alarm' occurs when a 'yes' response is elicited despite the S+ being absent, while responding 'no' even though the S+ was present is called a 'miss' (Macmillan, 2002). Derived from these, 'sensitivity', or hit rate, is the percentage of correct 'yes' responses for S+s, and 'specificity', or correct rejection rate, is the analogue of sensitivity for 'no' responses in the presence of noise alone (Repperger, Aleva, Thomas, Miller, & Fullenkamp, 2007). In addition, positive predictive value (PPV) and negative predictive value (NPV) are measures given by SDT analysis: PPV is the proportion of hits across all 'yes' responses, whereas dividing the amount of correct rejections across all 'no' responses yields a measure of NPV (Poling et al., 2011b).

From these data, the relative signal strength influencing behaviour (i.e., compared to the 'noise') can be calculated mathematically using the SDT

parameter of discriminability, or *d'* (Abdi, 2007), which is independent of the decision criterion. In general, a higher *d'* value would suggest higher perceived signal strength, associated with a lower proportion of incorrect responses over all responses (i.e., less 'misses' and 'false alarms', and more accurate performance; Kamil, Lindstrom, & Peters, 1985). In addition to these, SDT also provides a useful measure of bias, or β, in that bias towards one response alternative can be analysed separately from sensitivity to stimulus condition (Blough, 2001), giving an indication of the response strategy used by an organism (Abdi, 2007). That is, this measure indicates a tendency to respond either 'yes' or 'no' independent of stimulus discriminability, with accompanying changes in responses showing an inverse relationship between the incorrect responses (i.e. a bias towards 'no' responses would lead to a corresponding increase in 'misses,' but a decrease in 'false alarms;' Kamil, Yoerg, & Clements, 1988).

This theory is applicable across several scientific fields to a wide scope of problems and situations, both applied and experimental. SDT was first applied in psychophysics with radar studies and submarine detection (Abdi, 2007; Robin & McNeil, 1994) and has much relevance to human behaviour, but it is also applicable to many natural world settings, such as prey detection (Getty, Kamil, & Real, 1987). SDT continues to be applied in experimental settings and in studies exploring further theoretical explanations and applications of the way that important behavioural variables or methodologies interrelate with this framework (e.g., Blough, 2001).

Literature Review

SDT has its origins in the mid-twentieth century, with seminal works such as Green and Swets' (1966) *Signal detection theory and psychophysics* presenting the theoretical overview and examples of data and analysis for

practitioners utilising SDT. While the scope of this book is too broad to be examined here, a review of pertinent, current literature on SDT as well as relevant historical writings will follow.

A notable early example of both the theoretical underpinnings and experimental application of SDT is Tanner and Swets' (1954) article on visual detection. The authors conducted experiments in human visual detection using different light intensities as stimuli. Presenting the new theory of signal detection, a SDT analysis was then performed on the visual detection data, using mathematical and graphical evidence to contrast the SDT results against other models of the time (e.g., threshold theories). The results showed that there was no disadvantage in using SDT compared to other methods. The theory was found to be mathematically sound in retroactive analyses of much reported data, as well as offering advantages in terms of graphically reporting results for forced-choice and yes-no procedures; the internal consistency of the theory was demonstrated across both of these procedures. The authors emphasised that SDT addresses the element of decision-making in perception and detection, showing the need for psychological consideration of what might previously have been relegated to a physiological phenomenon.

Due to the focus on decision-making, many of these early studies utilising SDT had human participants only. In the early days of SDT, it was not entirely certain whether the principles and methodology would generalise to research with animals. In an attempt to generalise SDT across species, Nevin performed an analysis of data from a prior experiment with pigeons, where some unusual data related to response bias had caused analytical problems for the original researchers (Boneau, Holland, & Baker, 1965, as cited in Nevin, 1965). Boneau et al. were studying the effect of rewards on discrimination performance

between different light wavelengths, with responses to some stimuli leading to intermittent reward delivery; following reward delivery, on subsequent trials, a temporary increased tendency to respond towards even non-reinforcement-contingent stimuli was observed. However, the overall differential discriminative ability of the pigeons to the different stimuli was hardly negatively affected. In Nevin's (1965) graphical SDT analysis using a receiver-operating-characteristic (ROC) curve, the author was able to explain the pigeons' performance using the parameters of the sensitivity index (i.e., *d'*, or discriminability, which is relatively unchanging and a function of the stimuli), in comparison with the change-sensitive decision criterion, which, when lowered, increases the tendency to respond as if the S+ was present.

Despite this apparently satisfactory explanation, Nevin (1965) highlighted several areas of study needed in order to confirm that SDT was applicable to operant research with animals, including the effect of reinforcement, response effort, and unrewarded responding. However, following this paper, Boneau and Cole (1967) co-authored a comprehensive theory of SDT as it applied to animal discrimination behaviour. It was concluded that SDT provides an acceptable framework for this type of experimental investigation. It was posited that animal participants in SDT experiments work to maximise reinforcement. That is, repeated exposure to reinforcement contingencies which are associated with differentiated stimuli should theoretically allow for behavioural performance that matches reinforcement payoff values, which should result in comparable performance accuracy to even their ideally-informed human counterparts, despite strategic differences (Boneau & Cole, 1967). SDT research with animals and operant discrimination tasks is now well-established, as Nevin's (1965) identified

limitations have been adequately addressed in subsequent literature; the most relevant of those papers will be reviewed presently.

Nevin (1969) reviewed Green and Swet's (1966) book from the standpoint of how an operant researcher might utilise SDT. This analysis emphasised the role of concurrent schedules of reinforcement (i.e., two alternative schedules of reinforcement which are available at the same time, which an organism can freely distribute behaviour between; e.g. Herrnstein, 1958) and associated contingencies in typical yes-no experiments, as well as stimulus presentation probability and signal strength, as other variables that may have impacted performance. Of particular relevance to the present research is Nevin's (1969) description of matching performance under a SDT paradigm in a yes-no experiment where stimulus and reinforcement probabilities were varied.

Matching performance refers to the strict matching law (SML), a mathematical expression of behaviour which states that response rate for each alternative will be directly proportional to the rate of reinforcement for each alternative in concurrent schedules (Herrnstein, 1961). The probability of responding 'yes', regardless of the stimulus condition (i.e., presence vs. absence of S+) approximately equalled both reinforcement frequency and amount delivered for responding 'yes,' relative to that for 'no' responses; these results match operant studies of concurrent schedules (Nevin, 1969). However, the author suggested that signal strength (i.e., and the corresponding difficulty of the task) would play a mediating factor in this performance. With intense signals (i.e., an easier discrimination), accurate responses based on stimuli presentation would be expected, but for harder discriminations (i.e., towards zero signal strength), it was suggested that performance would instead simply maximise reinforcement. It was concluded that signal detection theory and operant conditioning can be suitably

integrated in terms of methodology, especially considering the advantages provided by the SDT measures of sensitivity and bias, which account for some of the practical problems encountered using traditional threshold theories in operant discrimination tasks.

Further integration of SDT and operant psychology was advanced through the work of Davison and Tustin (1978). Though traditional matching performance under a SDT framework was indicated in Nevin's (1969) review, the development of the generalised matching law (GML; see Baum, 1974) occasioned the quantified investigation of SDT performance as it relates to the GML for decision making in concurrent schedules (Davison & Tustin, 1978). The GML differs from the SML in that it has parameters which can account for undermatching (i.e., less extreme preferences than the SML would predict), overmatching (i.e., more extreme preferences than predicted by the SML), and bias (i.e., the degree of preference two one alternative over the other, even if both were equalised in terms of reinforcement). The authors' understanding of SDT yes-no procedures as operating under concurrent schedules echoed Nevin's (1969) earlier suggestions, and the authors applied this understanding to the SDT matrix presented in Table 1.

In a SDT yes-no task, there are two concurrent schedules involving both reinforcement and extinction. For the S+ condition, hits would be reinforced while misses were placed into extinction, and for the S- condition, correct rejections would be reinforced while false alarms would be under extinction, and the current reinforcement opportunities at any given time would be signalled by stimulus presentation (i.e. presence/absence of S+; Davison & Tustin, 1978; McCarthy, 1981). Behaviour under these schedules would follow the expected performance found with concurrent schedules: choice between alternatives would match their

relative reinforcement ratios. In SDT terms, this would mean that, varying directly with signal strength (Abdi, 2007; Nevin, 1969), if stimuli were readily discriminable, a response bias towards 'yes' in the presence of the S+, and a corresponding response bias towards 'no' in the absence of the S+ would be observed (i.e., hits and correct rejections, respectively; Davison & Tustin, 1978). Both the stimulus and reinforcement aspects of a SDT yes-no task were shown to agree with the principles of the GML, further integrating the psychophysical and operant psychological frameworks.

Following this paper, Davison and colleagues published a series of works investigating different variables and parameters of signal detection performance from a behavioural approach. In order to further clarify the effects of stimuli presentation versus the effects of reinforcement, McCarthy and Davison (1979) conducted three experiments with pigeons performing a yes-no task using light intensity of key-lights as the stimulus dimension of interest: the first varied both S+ (i.e., brighter light intensity, where S- was dimmer) presentation probability and reinforcement amount, the second varied S+ presentation probability while reinforcement amount across alternatives was equalised, and the third varied relative reinforcement across alternatives while the S+ presentation probabilities were held constant. It was found that behaviour varied as predicted by the SDT model of the GML for the first and third experiment, but not the second. This illustrated that the relative reinforcement ratio, rather than the S+ probability, was the salient controlling factor for SDT performance (McCarthy & Davison, 1979). S+ probability affected response ratios only inasmuch as it affected reinforcement ratios. In addition, the authors suggest that the results of these experiments could be affected by other, general biasing variables related to reinforcement in multiple and concurrent schedules. However, like Nevin (1969), the authors suggested that

discriminability as controlled by relative signal strength may have an additional effect along with those of the reinforcement parameters.

This view was echoed and extended by Lattal (1979) who examined the effect of using different types of reinforcement schedules as discriminative stimuli with pigeons. This experiment was arranged similarly to a traditional SDT yes-no task; however, the two stimulus alternatives (i.e. S+ and background 'noise', or S-) were 1-trial, 10-second schedules, both on a yellow centre key, requiring either the presence or absence of a key pecking response, respectively (i.e., differential-reinforcement-of-low-rates (DRL) and differential-reinforcement-of-other-behaviour (DRO) schedules; Lattal, 1979). For the majority of trials, there was then a 1.5-second consequential delay where either food or a blackout was presented for correct or incorrect responses, respectively, according to schedule type. After interacting with one of these schedules, the pigeons had to respond on additional associated keys, 'red' for the DRL schedule, and 'green' for the DRO. S+ (DRL) probability was varied. Incorrect responses resulted in a condition repetition until a correct response was observed, and all correct responses were reinforced.

It was found that, as DRL probability increased, the tendency to respond on the red key also increased (i.e., irrespective of response correctness). While response bias changed, discriminability between the reinforcement schedules was not systematically affected; however, removal of the 1.5 second delay between stimulus presentation and the choice alternatives increased sensitivity. The results suggest that schedules of reinforcement can serve as discriminative stimuli in SDT experiments. Similar to McCarthy and Davison (1979), the author concluded that, while there was a combined contribution of stimulus discriminability and reinforcement contingencies in SDT performance, response rate changes (i.e.

pecking vs. pausing) across the two schedule alternatives were more affected by variables like reinforcement frequency which have a direct effect on response bias (Lattal, 1979).

Later research by McCarthy and Davison (1981) once again identified reinforcement ratio as the chief predictor of response bias in SDT tasks, but also usefully analysed this response bias as having two sources: the reinforcement-related bias originating from differences across alternatives in amount, magnitude, quality, and latency, among other possible variables, as well as an inherent or constant bias occasioned by either the properties of the experimental task (e.g. one lever being harder to push) or by some qualities inherent to the participant of the experiment (e.g. preferring the colour of one key-light to another). In addition to acknowledging the role of reinforcement ratio, the effect of differences in the momentary value of reinforcers were also emphasised as contributing to the former type of bias (McCarthy & Davison, 1981). Thus, predicting experimental performance in an SDT task seems quite directly related to the GML and maximised reinforcement.

Although these examples of SDT research have produced some useful frameworks for additional theoretical development and analysis of decision-making behaviour, both experimental and applied, with humans and animals (Blough, 2001), issues have been identified in finding an applied, 'natural' situation that ideally fits these theoretical and experimental frameworks. While there are many applied examples where a SDT framework would be appropriate, variance in the consequences of signal detection in naturalistic settings with animals may bring the ecological validity of these experimental procedures into question. For example, in envisioning a real-world application of an SDT paradigm, it can be seen that correct rejections may not always be reinforced.

Consider the instance of an animal hunting for cryptic or mimetic prey; this prey must be detected against a wash of background information (i.e., 'noise'). However, correctly rejecting a non-prey object as unworthy of capture does not yield reinforcement as per the experimental situation, and the costs of false alarms and misses may not be equalised (Voss, McCarthy, and Davison, 1993).

Kamil and colleagues conducted a series of experiments on cryptic prey detection of the Catocala moth (*Noctuidae*) by the blue jay (*Cyanocitta cristata*) which followed these natural contingencies (see Kamil et al., 1985; Kamil et al., 1988; Pietrewicz & Kamil, 1977). In general terms, the experimental procedure followed the situation outlined above, where hits produced reinforcement, false alarms and misses either produced a cost (i.e. 30-second timeout from reinforcement) or had no consequences, and correct rejections had no consequences other than to occasion the start of the next experimental trial. For each trial, one slide either containing the moth (i.e., S+) or not (S-) was presented, with the response alternatives for the blue jays being to peck either an 'attack' key (i.e., analogous to a yes response in an SDT task) or a 'give up' key (i.e., analogous to a no response); this procedure can be termed a 'go/no-go' task. While it was found that the blue jays did not exhibit performance equivalent to that of a fully informed participant, in general, the birds did behave in such a way as to approach maximum reinforcement (Kamil et al., 1985; Kamil et al., 1988).

Voss et al. (1993) investigated SDT task performance under similar contingencies using two experiments. The first procedure presented a standard yes-no SDT task where all correct responses were reinforced. For the second experiment, the non-reinforcement of correct rejections, where only hits were reinforced, was compared to their first procedure. Using two light intensities on a centre key as stimuli, six pigeons were trained to respond on a left key following

the brighter (i.e., S+) initial key presentation, and on a right key if the centre key was initially dimmer (i.e., the S-). For the second procedure, hits led to three seconds of reinforcement, while correct rejections and misses led to 3-second periods of timeout where reinforcement was unavailable. For both procedures, the timeout durations for false alarms were varied between 3-120 seconds with respect to the ecological variance of the different responses and their associated costs.

Detection performance varied systematically with changes in this timeout duration for the first procedure, but had inconsistent, unsystematic and idiosyncratic results in the second procedure. In general, accuracy was higher in the second procedure. A large bias towards 'yes' responses was observed, due to that being the only response that provided reinforcement. This bias differs from the results of Kamil et al. (1985), but the authors suggest that this is due to the response requirement differences across the response matrix between the studies (i.e. the 'costs' associated with each response type). In order to respond 'yes,' the jays had to complete a fixed-interval 30-second schedule (FI30), but there was only a single response requirement with no waiting-time requirement for 'no' responses; in the study by Voss et al. (1993), response requirements were equal for 'yes' and 'no' responses. They acknowledge the additional biasing factor that differential response requirements have upon SDT performance. In conclusion, the authors suggested that evaluating SDT performance using these naturalistic contingencies is a robust methodology; the principles of operant psychology and SDT can still be integrated even when correct rejections are not reinforced.

It has been demonstrated that SDT is appropriate for a behavioural analysis of detection and decision-making behaviour both theoretically and experimentally; the utility of this framework for applied research in the present

day will now be evaluated. As previously mentioned, SDT was initially developed in applied studies of visual detection with radar operators, such as some of the initial research of Green and Swets (1966). The application of SDT to human decision-making has provided useful understandings and solutions to problems in many different branches of the field of psychology, as well as other scientific disciplines, such as: the clinical practice of psychologists in terms of evaluating the correct detection of the mistreatment of children, predicting risk of disorder relapse, and evaluating treatment response, among others (McFall & Treat, 1999); improving human decision-making in terms of the risk judgments made by health and social professionals evaluating financial elder abuse in terms of harm reduction and formulating better training interventions (Harries et al., 2014); understanding the organisational decisions of entire governmental systems, such as the child welfare services of the United States of America (Mumpower & McClelland, 2014). It has been demonstrated that SDT can be usefully applied in studies of human decision-making related to topics of important social consequence. In addition, as Nevin (1965) hoped, SDT has also been usefully applied to animal behaviour, providing solutions to important practical problems of the human and animal experience. This is exemplified in research by Poling and associates, who trained giant African pouched rats (*Cricetomys gambianus)* in detection tasks of relevance to safety and health initiatives.

Following a procedure similar to the naturalistic contingencies used in the Voss et al. (1993) study, Poling, Weetjens, Cox, Beyene and Sully (2010) trained these rats in a scent detection task to signal the position of buried landmines. Using the principles of operant conditioning, the rats were systematically and gradually trained from being able to detect a common explosive ingredient (i.e., variants of trinitrotoluene, or TNT) in sand samples, to

being able to search for and detect buried, defused landmines in increasingly larger plots of land. The final testing criterion was a perfect hit rate (i.e., all mines were correctly identified) with less than two false alarms in a 100-m^2 area on a blind test run, where the rat's handlers did not know the location of any of the mines. Indicator responses that occurred further than 1 m from a mine were recorded as false alarms.

Throughout training, reinforcement was delivered contingent only on hit responses; false alarms, correct rejections, and misses were not reinforced, and misses were occasionally subject to a correction procedure of repeated stimulus presentation at the discretion of the handlers. Following accreditation as appropriate demining animals, the rats were sent for fieldwork in clearing actual mines, where they continued to be subject to both training and testing; the operational criterion was a perfect hit rate and less than 5% false alarm responses. In the field, due to the handlers' uncertainty in differentiating between S+s and S- at the time of indication response, no responses were reinforced. Correct responding was maintained by intermittent reinforcement of hits in the simulated minefield training areas. This article illustrated how SDT tasks can be applied to humanitarian issues and animal behaviour in the field, also demonstrating that highly accurate performance can be elicited using the naturalistic contingencies described by Voss et al. (1993). A further paper by Poling et al. (2011a) evaluated the rats' performance in the field and demonstrated the efficacy and real-world advantages of this SDT-based operant training.

In addition to landmine scent detection, further work with these rats indicated potential application of a similar training paradigm in order to detect the presence of tuberculosis (TB) in human sputum samples (Poling et al., 2011b; Reither et al., 2015). Though TB infection is prevalent and a leading cause of

death in patients with HIV in sub-Saharan Africa, there was a distinct need for the development of an economical test alternative to microscopy (i.e., the most common testing method) that also improved the detection of positive samples above historically problematic levels. The researchers attempted to develop an economical and quick diagnostic test having both high sensitivity and specificity by using the rats as detectors: pausing by a sample would constitute a 'yes' response, and passing over a sample would be a 'no' response. Procedurally, this is an applied example of the go/no-go procedure, like that used by Kamil (1985), similar to the procedure used in this study. While initial efforts suggested that this would be an economical, rapid and accurate method of testing (Poling et al., 2011b), later findings suggested that, although the rats provided fair-to-moderate accuracy across both HIV-infected and uninfected patients when compared to the gold standard of testing, bacterial culturing, their performance was linked to bacteria count, with higher concentrations being more easily discriminable (Reither et al., 2015). It was concluded that the rats alone do not provide satisfactory diagnostic performance to serve as standalone testers in high-endemic settings at this time, with regard to the guidelines of the World Health Organisation. However, there is a suggested application for this work as a second-line screening methodology, as the rats can inexpensively handle a high-volume caseload.

Thus, it can be seen that SDT provides a sound theoretical framework for the experimental understanding of detection, discrimination and decision making behaviour with both humans and animals. SDT also has provision for application in important research with humans and animals which can provide solutions to pressing humanitarian issues of our time, including harm reduction and the treatment of disease. The effect of reinforcement contingencies on a typical yes-

no SDT task can be expected to follow performance predicted by the generalised matching law in cases where both alternatives are reinforced, but this procedure is also robust when there is only one source of reinforcement between decision alternatives. Thus, in referring to Nevin's (1965) comments on the potential limitation of SDT applications, the roles of both reinforcement and unrewarded responding on SDT performance have been well documented in the intermediary literature. However, it appears that the role of response effort has not been investigated as thoroughly.

Effort

The role of effort as a variable invites important consideration in the analysis of behaviour. In a broad sense, effort can be defined as "the subjective sense of the amount of work required to perform some voluntary action" (Reed, 2010, p. 799). In practical terms, effort has been equated to several different paradigms, all of which vary some dimension of a task to increase this amount of 'work' required, including physical effort (i.e., calories expended; Horner & Day, 1991), task difficulty (Reed & Martens, 2008), force requirements (e.g. Alling & Poling, 1995; Chung, 1965), and ratio requirements (Elsmore, 1971; Zentall, 2013), with differential effects observed upon both human and animal behaviour when response effort was varied. Generally, findings exploring the role of effort in the study of behaviour have shown that effort exerts some control over choice behaviour, with less effortful behaviours being performed more frequently than their more effortful counterparts when other variables are equalised; this has come to be known as the 'least effort principle' (Reed, 2010). There is also evidence that increasing the amount of effort required to earn a reward can decrease the frequency of the associated response (Nishiyama, 2014). However, in contrast to expected performance, an 'effort justification' effect has also been observed in

both humans and animals, whereby rewards received after a high amount of effort have a greater value placed upon them (Lydall, Gilmour, & Dwyer, 2010). The literature will now be reviewed in detail to assess the role of effort as a variable in behavioural processes.

Literature Review

Across the operant psychology literature, the general conclusion related to response effort is that this variable, in combination with the effects of reinforcement, influences behaviour as per the GML. When reinforcement across alternatives is equalised, the less effortful option is preferred (Wilson, Glassford, & Koerkenmeier, 2016). With reference to the GML (Baum, 1974), this least effort principle can be explained by response effort generally impacting behaviour as a variable affecting response bias (Reed & Martens, 2008). That is, more effort required by one alternative has been found to increase bias towards other, less effortful alternatives, suggesting that increasing effort acts aversively (i.e., as a punishing factor which reduces the frequency of behaviour), independent of and in addition to those of reinforcement, although recent research has called this into question, suggesting that, in some cases, this could better be explained as a response measurement issue (i.e., as force increases, 'responses' below force effort criterions may go unrecorded; Pinkston & Libman, 2017). In addition, increasing response effort has been shown to decrease overall response rates, and vice versa (Chung, 1965; Friman & Poling, 1995), as well as total response amounts, and to increase the time between responses, post-reinforcement pauses, and response latency (Alling & Poling, 1995; Elsmore, 1971). While this effect could be a function of the increased time necessitated by an additional effort requirement, these findings could also support the suggestion that increasing effort is aversive.

The aversive effects of response effort have been examined as controlling factors in the management of problem or undesired behaviour, especially in cases where reinforcement contingencies are difficult to control. Horner and Day (1991) manipulated the efficiency of functionally equivalent alternative responses to aggressive problem behaviours with humans, varying the parameters of physical effort, reinforcement ratio, and delay to reinforcement in three separate cases. They found that alternative, reinforcement-contingent responses that were less efficient than the baseline aggressive responses on any of those parameters were not effective behavioural alternatives. However, when the responses were made more efficient (i.e., effort was decreased), they were able to effectively compete with the problem behaviour.

The treatment of self-injurious behaviour (SIB) in humans that is maintained by automatic reinforcement is another area where manipulating response effort allows for more parsimonious behavioural control than attempting reinforcement contingency control. Although Zhou, Goff, and Iwata (2000) acknowledged that it was unclear whether decreasing reinforcement or applying punishment caused the relevant effects in their intervention results, an inhibitory effect of increasing response effort (i.e., force required to perform the behaviour) was observed with hand-mouthing SIB, and a corresponding increase in object manipulation (i.e., the other behavioural alternative) was demonstrated across all participants. These results reversed correspondingly when baseline conditions were reintroduced and again when the effort condition was repeated. The findings had both clinical and applied relevance, suggesting that the increased effort was the salient variable affecting behaviour change, despite the fact that automatic (i.e., non-social) reinforcement for SIB was always available. The authors concluded that increasing the response effort for SIB can be an alternative means

of decreasing this problem behaviour and suggested further application to other problem behaviours.

This inhibitory effect of effort was also found with the automatically reinforced problem behaviour of non-nutritive object ingestion, or pica, in humans (Piazza, Roane, Keeney, Boney, & Abt, 2002). More comprehensively, this study also examined the effect of response effort on the pica behaviour alone, and when effort was systematically varied for both the SIB and other behaviours of alternative, preferred item acquisition. In the absence of alternative items (i.e., the pica behaviour alone was examined), increasing response effort decreased observed levels of pica in relation to baseline measures. In addition, when both behavioural alternatives had equal effortful 'costs,' behaviour leading to access to the alternative items was increasingly performed over that for pica items, and levels of pica were consistently reduced in relation to baseline when given the opportunity to gain alternative items, even when doing so required a high response effort. Relative to baseline, even when the cost of the alternative behaviour was high, levels of pica responding were decreased.

Overall, these findings supported the indications of the GML with regard to the biasing effect of effort, and that the effects of variance in effort for both response alternatives interacted with those of reinforcement, specifically reinforcement quality. As effort for one alternative increased, the other response was increasingly performed, with differential effects depending on the level of effort required between alternatives. For example, if there was already a high cost for pica, increasing the response effort for obtaining the preferred alternative items resulted in little to no increase in pica-related behaviour. It can be seen that the findings of this study, specifically as they relate to the relative costs of responding to two alternatives, can be theoretically applied to understanding

differences in the results of Kamil et al. (1985) and Voss et al. (1993), further substantiating the authors' suggestions about the differential impact of response effort across response alternatives.

Friman and Poling (1995) provided a review of response effort as a variable in behavioural studies, offering additional support for this aversive or inhibitory effect of effort. The authors demonstrated that, while all behaviour is economically weighed in terms of relative costs and benefits, increased effort contributed to this as a cost, effectively reducing response rates even across different components of chain schedules, and in extinction. Behaviour having extreme effort requirements has even been shown to elicit behaviour that functions to obtain escape and avoidance of contexts associated with the effortful responding. The collected findings illustrated that the long-term effects of effort on behaviour are similar, although distinct from, several characteristics of punishment. The authors demonstrated the relevance of response effort for both experimental behavioural studies and applied work, illustrating the effective use of increased effort for reducing problem behaviours, as shown above, but also how decreased effort could increase rates and probability of eliciting prosocial or desirable behaviours.

While several examples of effort relating to human behaviour have been given, the relative effort required has also been shown to bias animals' stimulus preference in choice situations, such that increasing response costs (i.e., the fixed-ratio, or FR, requirements) decreased previously established preferences related to discriminative stimuli, and reversal of these preferences varied systematically with changes in FR requirements (Roper & Zentall, 1999). This biasing factor was found to be an additive biasing factor of overall preference in conjunction with the probability of reinforcement. In addition, it has been observed that, when more

effort is required in the form of a larger FR requirement, the choice responses made by animals are more accurate. Rohles (1961) conducted an experiment using visual stimuli where an "odd" stimulus must be differentiated from two "like" stimuli in a three-stimuli array; correct indications in stimulus selection were initially reinforced under a continuous reinforcement (i.e., CRF, or FR1) schedule, but as the ratio requirement was increased, choice behaviour increased in accuracy. However, as the order of responses was not varied across the array, it is possible that this effect on accuracy was instead a product of a learned sequence of responses. Finding a similar effect on accuracy in a delayed matching to sample experiment, Spetch and Treit (1986) conceptualised it as the animal avoiding making errors after performing a larger amount of work, due to the cost of such errors, but also as a function of allowing increased exposure to the stimuli associated with choice alternatives, rather than due to the expense of effort. Despite this, there has been some evidence for a preference toward tasks requiring higher effort when these are associated with higher reinforcement rates and quality relative to less effortful tasks (e.g., Lydall et al., 2010; Neef, Shade, & Miller 1994, as cited in Billington & DiTommaso, 2003).

For example, Clement, Feltus, Kaiser, and Zentall (2000) investigated the finding that stimuli that follow increased effort or longer delays were preferred over their less effortful or sooner counterparts. Pigeons were trained in two simultaneous discrimination tasks, where different schedules were presented with the initial stimulus before presenting the discrimination choice between two additional stimuli; one task was a FR1 for both S+ and S-, the other was a FR20 for a different S+, S- pairing. Following this training, when presented with a choice between stimuli, the pigeons consistently preferred the stimuli that had followed the FR20 schedules in training.

Further research by Lydall et al. (2010) clarified the differential effects of effort versus delay, as well as the role of reinforcer value, in observations of this effort justification effect. Using rats, a lever-pressing task was divided into high- and low-effort conditions (i.e., based on FR requirement), and reinforcer value was measured using a previously established empirical measure of the clusters of licking behaviour for a liquid sucrose reinforcer. Reinforcement delivery to additional rats who did not have access to levers was yoked to performance of rats in the master conditions to measure the effects of delay to reinforcement relative to that of effort. All rats placed a higher value on reinforcement obtained in the high-effort condition, and this difference was greater for the rats who were under the schedule requirements (i.e. the master FR conditions requiring lever pressing), suggesting, while time to reinforcement affects preference, that this delay has a less impactful effect than effort itself.

While this performance appears to violate the prior conclusions regarding the effect of effort on behaviour, an explanation was offered for this phenomenon in the effort justification hypothesis, which emphasised the justifying effect of subsequent reinforcement as a process related to cognitive dissonance (Aronson & Mills, 1959, as cited in Zentall, 2013). While this theory was developed to explain human behaviour and relied heavily on cognitive assumptions, a similar effect was observed with animal behaviour, necessitating a more parsimonious explanation of this phenomenon. Zentall and colleagues conducted a series of experiments with animals systematically examining this effect, with results consistently demonstrating that the consequences following aversive events (i.e., including effort as measured by force and reinforcement schedule requirements, amongst others) were consistently preferred over the products of less aversive events (see Zentall, 2013).

Thus, the within-trial contrast (WTC) hypothesis was formulated: the relative aversiveness of the effort requirement can serve to increase the value of a subsequent reward because of the hedonistic contrast it provides (Zentall, 2013). Again, effort is indicated as a biasing variable, but it is suggested that, in some situations, increased effort associated with a choice option may bias behaviour in favour of that alternative. However, these effects were not always found to be replicable. Tsukamoto, Kohara, and Takeuchi (2017) suggested that this was due to idiosyncratic differences in the experience of aversiveness produced by effort leading to the WTC effect; if participants failed to perform an amount of effort that was sufficient to experience aversiveness, then the relative change from aversion to pleasure (i.e., as assumed from reinforcement) would be decreased, and the overall effect would be reduced. Similar to the findings of Lydall et al. (2010), Tsukamoto et al. (2017) observed a preference for events following greater effort in humans, but not for those following greater delay to reinforcement. Thus, the evidence appears to suggest that behaviour associated with increased effort is less preferred, but stimuli and reinforcers following increased effort expenditure are more preferred, although further research into the WTC is warranted by the apparent lack of universal applicability of its principles.

However, this could also be due to differential effects of different types of 'effort' being falsely equated over studies. Early research comparing two common measures of effort, force requirements and number of response requirements, indicated that FR requirements resulted in a more stable effect on behaviour in terms of resistance to extinction than force requirements, although this effect was not always linear (Weiss, 1961, as cited in Gonzalez, Bainbridge, & Bitterman, 1966). Elsmore (1971) also compared the effects of force versus FR requirement in a discrimination task where the two stimuli had different

reinforcement probabilities. Pigeons pecked a key illuminated with either red or white light, having an associated reinforcement probability of .25 and .50, respectively, and nonresponding terminated a trial after eight seconds. In two experiments, every 10 sessions, first force (i.e., between 25-150 grams) and then FR requirements were systematically increased across trials, up to a maximum of FR64, before being returned to a FR1 for five sessions.

Using force, idiosyncratic effects upon performance were observed across the pigeons; in general, behaviour was more variable under the lower reinforcement probability condition, and only relatively high force requirements resulted in differential responding for some birds. Using FR requirement, while differences were minimal at low FRs (e.g. FR1, FR4), increasing FR value caused an increased difference in performance between the two stimuli. Responding on the white key showed very little variance overall, except at the very high FR64. In contrast, increasing FR for the red key resulted in decreasing percentage of responses overall, decreasing response rates, and increasing response latency; this supported the findings of Chung (1965) where increased effort resulted in decreased response rates, but generalised this to another modality of effort, FR requirement. A comparison between experiments illustrated the relative advantage of using FR requirement over force as a measure of effort, because it was less affected by idiosyncratic responding, producing more consistent results across the pigeons. The author concluded that response effort is an important variable to consider in studies of discrimination and stimulus control.

Research Aims

The literature clearly suggests that response effort exerts control over behaviour; however, at present, it is unclear how mechanisms of this control factor into a sensory detection paradigm, especially one that attempts to replicate

more natural contingencies, such as those used by Voss et al. (1993). I was unable to discover any research exploring the effect of systematic variations in effort as measured by reinforcement schedule requirements assessed using SDT where only hit responses were reinforced. The present research sought to address this, utilising a procedural variation on the standard SDT methodology derived from the studies by Kamil et al. (1985) and Voss et al. (1993), similar to that of the applied work of Poling and colleagues (Poling et al., 2010; Poling et al., 2011a; Poling et al., 2011b). This SDT paradigm has been utilised in a range of studies with avian species, and the present study extended this to work with domestic hens (*Gallus gallus domesticus*), using key-light stimuli that differ in brightness levels, similar to the stimuli used by Voss et al. (1993); effort was systematically varied using FR requirements.

It was hoped that this systematic experimental investigation may yield important information about the role of effort in an SDT paradigm which may be of use to support and inform applied research in this field (e.g. the work of Poling and colleagues). As response effort will be necessarily included to some degree in every SDT task, the utility of this research was promising with regard to interpreting past findings from SDT studies, both experimental and applied, but also in informing future research. It was hypothesised that, due to relative reinforcement, there was likely to be a bias towards 'yes' responses, especially at lower FR requirements. However, as response effort increased, it was theorised that sensitivity (i.e., hit rate) would decrease, while specificity (i.e., correct rejection rate) would increase. The linearity of this relation was unknown, and expected behaviour at extremely low or high FR requirements was unknown.

Method

Subjects

Subjects were six domestic hens (*Gallus gallus domesticus*), nominally categorised as 12.1 through to 12.6. At the beginning of the research, all hens were two to three years old. All of the hens had prior experimental experience, including some key-pecking (i.e., 12.2, 12.3, 12.5, and 12.6) and screen-pecking (i.e., 12.4) experience; 12.1 was the only hen experimentally naïve with respect to pecking requirements.

For the duration of the research (i.e., approximately 13 months) hens were housed individually in wire cages, measuring approximately 500-mm long by 510-mm wide by 390-mm high. In the windowless, ventilated colony room, cages were arranged six to a row, side-by-side, with three rows on the left and right sides of the room; over the course of the experiment, the total number of chickens in the room was between 6 and 36. The light-dark cycle of the room was controlled, maintaining 12 hours each of light and dark at starting from 6am. In the home cages, hens were given ad libitum access to water, as well as any necessary supplemental post-experiment feed (commercial laying pellets). In addition, once per week, hens received supplemental vitamins as needed and grit to promote overall health. To address any illness or injury, as needed, hens were administered an oral non-steroidal anti-inflammatory and/or a broad-spectrum antibiotic under the advice of the laboratory veterinarian and the supervision of the chief laboratory technician.

To monitor ongoing health, hens were weighed prior to every experimental session, and weighed approximately every second laboratory running day if a hen was not participating in experimental sessions for any reason (e.g. illness, injury, conditions being met). Before the experimental sessions

commenced, hens' free-feeding bodyweight was determined over 10 days of ad libitum access to feed and water where no trend in weight fluctuation was observed using daily weighing. For the duration of the study, hens were reduced to a lower bodyweight, with their reduced bodyweight being maintained at approximately 80% or greater relative to free-feeding bodyweight. Ethical approval for this study was given by the University of Waikato's Animal Ethics Committee (Protocol 986), and the practices of care complied with the relevant Standard Operating Procedure registered with this committee (#17 Care of Poultry During Behavioural Research), the University of Waikato Code of Ethical Conduct for the Use of Animals in Teaching and Research (2014 version), and the New Zealand *Animal Welfare Act* (1999).

Apparatus

This experiment utilised a ventilated plywood chamber box, measuring 565 mm long by 400 mm deep by 532 mm high internally. Access to the chamber was through a side panel having a hinge on the left-hand side; when this door was closed, the interior of the chamber was darkened, as there were no window panels on the box. All internal walls were bare and painted plain white, aside from the right-hand wall, where the key-lights and magazine feeder were situated, as shown in Figure 1.

The two, translucent key-lights were 30 mm in diameter, positioned within metal surrounding plates (measuring 70-mm wide and 140-mm high, with 87 mm between them), located approximately 350 mm from the chamber floor, requiring a relative force of ≥0.2 N to record a response. This force was kept constant throughout the experimental procedure. When on, the left key-light was transilluminated with a green hue, while the right key was red (see Figure 1).

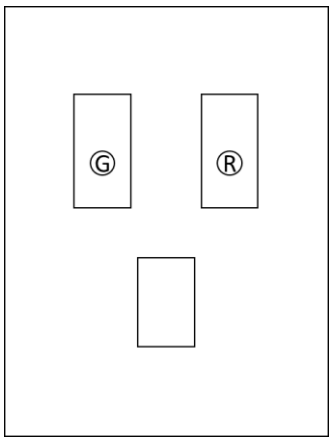

Figure 1. Schematic of the right-hand, internal wall of the chamber box, illustrating the locations and relative size of key-lights and magazine feeder access (i.e., the lower box).

A feedback beep was played upon every effective response to a lit key-light; responses to unlit keys did not result in any programmed consequences or the feedback beep. The key-lights were connected to programmable light-emitting diodes (LEDs), allowing them to display different colours and levels of brightness, using pulse-with-modulation (PWM) programming to vary the latter. The magazine access consisted of a hole that was 70mm wide and 110 mm high, allowing the hens to place their head into the wheat delivery portion of the magazine. The magazine was lowered (i.e., out of reach of the hens) until the reinforcement cycle, at which time the food hopper was raised to allow access to the wheat. The experimental procedure, including delivery of reinforcement and data collection, was automated using Med-PC® IV software running on a Dell computer, but data were also recorded by hand into a ledger at the end of every session.

Procedure

Advance key procedure. This experiment utilised a procedure where all stimuli used in the discrimination task were presented on the left-hand, green key-light. At the beginning of a trial, this key-light would transilluminate with either a

brighter stimulus (S+), which varied across hens (see Training: Phase 4), or a dimmer stimulus (S-), which was held at a constant brightness level across all experimental sessions and subjects. Key-light LED brightness was controlled through software utilising PWM, where a coded value of 0 would result in a signal strength of zero (i.e., the light would turn off), and a value of 255 would result in the LEDs being fully illuminated; that is, programming 0 resulted in 0% brightness, and programming 255 resulted in 100% brightness, with mathematically proportional changes across all values in between (see Hirzel, 2017, for additional explanation). Stimulus presentation probabilities were 50% for both the S+ and S-, and the presentation of these stimuli were randomised.

When either the S+ or S- was presented on the green key, one observation response was required to progress to the indication, or discrimination, phase; the right key remained unlit. When the hen performed the required FR 1 key peck, the beginning of the indication phase was signalled by the illumination of the red key. At this point, both keys were illuminated. If the S+ was present, responses on the left key that fulfilled the current FR requirement (i.e., 'hits') would lead to reinforcement; hit rate (HR) was calculated as the proportion of S+ trials on which this FR requirement was fulfilled over all S+ trials. If the S- was present, responses on the left, green key would have no effect. For both trial conditions, a single key peck on the right, red, 'advance' key had no consequences other than to advance the procedure to the next trial regardless of the actual or presence (i.e., miss) or absence (i.e., correct rejection) of the S+; correct rejection rate (CRR) was calculated as the proportion of S- trials ended with the red key (i.e., prior to reaching the FR requirement for S+ trials by pecking on the green key) over all S- trials. Thus, responses to the green key were analogous to 'yes' responses in a standard yes-no SDT discrimination task, while responses on the

red key corresponded with 'no' responses. Responses on the green key in the absence of the S+ had no programmed consequences, as per naturalistic contingencies. Whether the trial was ended by fulfilling the S+ FR, or by pressing the advance key, there was an inter-trial interval (ITI) of two seconds where both key-lights were turned off, before the next trial began with the left key presenting either the S+ or S-. Figure 2 graphically represents the different trial types described here.

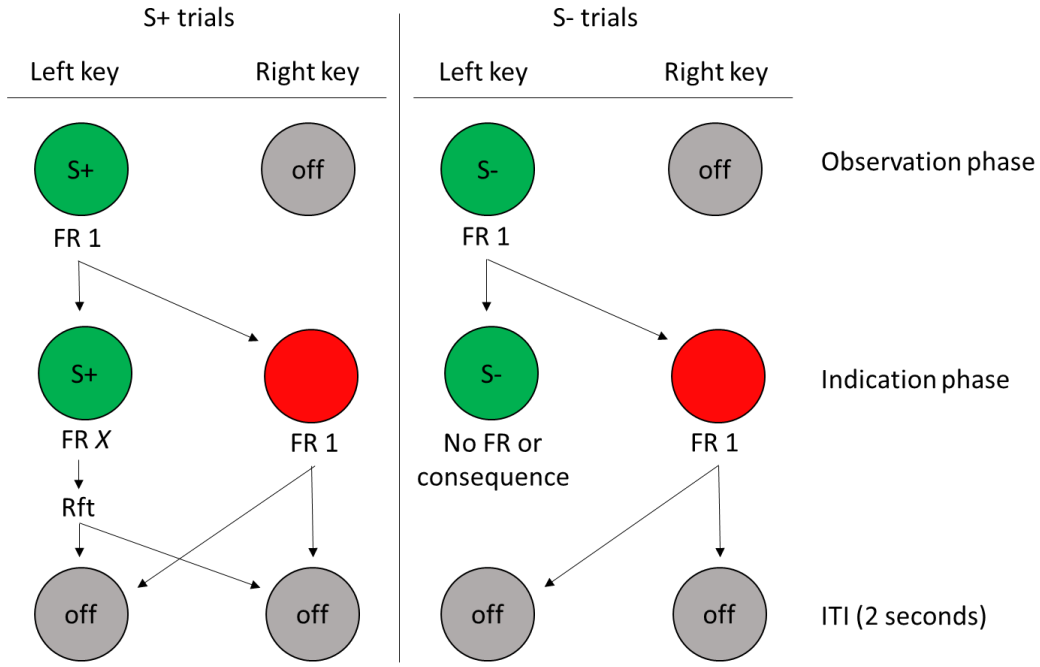

Figure 2. The three stages in each trial of the advance key procedure for S+ and S- trials. 'FR *X*' is used to denote the response requirements for the indication response. 'Rft' refers to reinforcement delivery.

Laboratory sessions. In general, one experimental session was run per day, between five and seven days a week, with a variety of personnel running the experiments, generally beginning between 8 and 9 am, but also across a wider range of different start times. Experimental sessions terminated after 40 reinforcers or 40 minutes, whichever criterion was reached first. Across all sessions and conditions, 3-second access to wheat was used for reinforcement.

Shaping. Due to the differences in the hens' prior experimental experience, all hens were initially shaped to peck either of the key lights using a CRF schedule; the magazine was operated by hand, and reinforcers were delivered for responses that were successively closer to the keys, and eventually for pecking the left-hand key. When several key-pecks had been observed (i.e., varying between 10-15 depending on the hen), the hens were judged ready to move on to the first phase of training at the researcher's discretion.

Training. Four phases of training were undergone by each hen. Across all training phases, and in the experimental sessions, the observation FR value was set at FR 1. Similarly, the advance key FR value (AdFR) was also set at FR 1, to keep this effort requirement constant across all trials and relatively low, corresponding with naturalistic detection behaviour in terms of the relative ease of a rejection response (e.g., a blue jay continuing to scan for prey instead of making an attack response).

Phase 1: Key peck training. To empirically establish that hens were ready to move on to the discrimination training, the hens' key-pecking performance was assessed using only S+ trials of the advance key procedure. S+ brightness was set at a value of 255 (i.e., 100%). For this initial training phase, the indication FR (IndFR) requirement was set at FR 2. Criterion for moving to the next training phase was to obtain all 40 available reinforcers before the session would terminate after 40 minutes; all hens passed this requirement on their first session.

Phase 2: Discrimination training. Once a reliable key-peck response from all hens had been established for the S+ alone and the criterion for phase 1 had been met, S- trials were also introduced to the session, with 50% randomised stimulus probability. S- brightness was set at a value of 10 (i.e., 3.92% of total

brightness available) to make the differences between S+ and S- stimuli readily discriminable. All FR values remained as in phase 1, and only hits were reinforced. Each hen was exposed to 23 sessions of discrimination training under these conditions, to ensure sufficient exposure to the contingencies of the task. The criterion for moving to phase three was performance consistently no less than 1.6 for the combined measure of HR+CRR; this value was selected as it lies between chance (i.e., binary guessing; 1.00) performance and perfect performance (2.00), slightly above the halfway point.

Phase 3: Increasing FR requirement. Following the discrimination learning, using the same procedure, the IndFR value was systematically increased stepwise by two to reach the desired baseline of FR 10. This baseline was selected as high accuracy had been observed under FR 10 in similar experiments in this laboratory (unpublished data). The criterion for increasing the FR was two consecutive sessions where no less than 1.6 on the HR+CRR measure was observed.

Phase 4: Decreasing brightness of S+. Once the FR 10 condition was reached, hens were given 18 sessions to establish consistent performance at the higher FR values before the stimuli brightness levels were manipulated. Due to the large, almost maximal difference in brightness between the S+ and S-, performance across all hens, as measured using hit rate and correct rejection rate, was consistently close to perfect (i.e., for the separate measures, performance generally ranging between 0.95 and 1.0, where 1.0 indicated 100% accuracy of discrimination). This meant that any changes in behaviour would most likely be obscured by ceiling effects when graphically analysing the data.

In order to more clearly illustrate behavioural changes due to FR variation, the brightness level of the S+ was systematically adjusted. As indicated

by previous research from this laboratory (unpublished data), hens were able to discriminate between PWM values with as little discrepancy as 3 integers (i.e. between 53 and 50 PWM values); thus, the S+ brightness PWM level was quickly decreased to 155, then 55, 30 and 20, with at least two sessions per condition change, until it was at a PWM value of 15 (i.e., 5.88% of total brightness, relative to the S- of 10 PWM and 3.92%).

Once at 15, the criterion for adjusting the S+ was relative to a performance band of 1.4-1.6 (i.e. inclusive) on the HR+CRR measure, for reasons outlined previously. If there were at least two consecutive sessions where performance was above this band, the S+ brightness was decreased by a PWM value of 1, with brightness being increased by 1 if there were at least two consecutive sessions below the band. Brightness was decreased overall, until performance was recorded for 5 consecutive sessions within the band.

Table 2

Stimuli Brightness Levels for Experimental Trials

Hen	S+ PWM Value	S+ Percentage of Total Brightness
12.1	12	4.71%
12.2	13	5.10%
12.3	13	5.10%
12.4	12	4.71%
12.5	13	5.10%
12.6	13	5.10%

Table 2 delineates the different brightness levels used for the S+, listing the PWM values and percentage of total brightness for the key-lights across all hens, as the levels were individually set.

The S+ levels for all hens were set without issue, with the exception of 12.4, whose optimum S+ value based on this criterion appeared to be a brightness level somewhere between the PWM values of 12 and 13; however, with this programming, it was only possible to set brightness values using positive integers. Thus, the lower S+ value of 12 was selected and set following 5 consecutive sessions where all but one data point was within the band. This point measured 1.39, and so the difference was considered minimal and acceptable to move to the experimental phase.

Experimental sessions. All experimental sessions included trials as described in the advance key procedure outlined above, utilising the individualised S+ values for the different hens as specified in phase four of training. As in all pre-experimental phases, the S- brightness level was held constant at a PWM value of 10 (i.e. 3.92%). From the baseline condition of FR 10, the FR requirements in the indication phase of every trial were systematically varied across sessions, increasing stepwise by three; however, due to human error, there was one condition (i.e. when increasing from FR 31) where the FR was only increased by two to a FR 33. From this point, incremental increases of three were reinstated. The criteria for increasing the FR requirement was at least 500 trials across a minimum of seven consecutive sessions, with data from the HR+CRR measure of the four most recent sessions demonstrating behavioural stability, as assessed by visual analysis to ensure the absence of a trend. The trial and sessions requirements in these criteria were to ensure that data were not influenced by sessions where the hens emitted little behaviour, or no relevant behaviours at all.

Termination criteria for ceasing to increment the FR requirements was having fewer than 5 reinforcements delivered per session for four consecutive sessions.

Baseline (i.e. FR 10) was reinstated when hens had reached the termination criteria. However, due to time constraints associated with the laboratory being closed down, only one hen (i.e., 12.4) met termination criteria. Baseline was reinstated for all other hens at the same time to allow adequate time for previous conditions to be repeated before the laboratory facilities became unavailable. Following baseline, the initial probe condition repeated the median FR requirement between the baseline condition (FR 10) and the highest FR requirement the hens had worked under (i.e., median probe condition); this varied individually. Where the median was between a pair of conditions, the condition with the lower FR requirement was substituted. The next probe condition was the median between the median probe trials and the maximum FR (i.e., three-quarter probe condition). Following this, FR 13 requirements were reinstated for all hens. For 12.4, probe trials were then conducted at levels lower than baseline (i.e., FR 8 and FR 5), and finally at the FR requirement immediately preceding the terminal FR. All reinstated conditions, including baseline, had a minimum of five sessions, with 12.4 having seven sessions for each condition, as time allowed. All sessions were conducted once per day, with the only exception being the penultimate session of the last condition for 12.6 which, due to mechanical failure, was conducted six hours later, on the same day as the preceding session.

Results

Figures 3, 4, 5, 6, 7, and 8 display the performance of individual hens as the FR requirement was systematically increased from the original baseline condition to the terminal condition that was in effect before the baseline condition was reinstated. This performance includes both specificity and sensitivity data, as well as a combined measure of both. In all graphs, specificity, or the proportion of correct 'no' responses to all trials where S- was present (i.e., pecking the red key prior to reaching the specified FR requirement on the green, 'yes' key), is represented by the red lines with circular data points; sensitivity, or the ratio of correct 'yes' responses for all S+ presentations (i.e., pecking the green key and completing the FR requirement in effect), is represented by the green lines with triangular data points; and the combined data from both of these measures (sensitivity + specificity) is represented by the black lines with square data points. Both specificity and sensitivity are ratio measures, having a range for possible values from 0.00 to 1.00; the combined measure ranges from possible values of 0.00 to 2.00.

Variables that were observed to impact hen performance were: malfunctions in experimental apparatus (i.e., both hardware and software), illness/injury (e.g. 12.6 developed an eye infection), and hen laying cycles (i.e., laying eggs in the chamber box); where possible, these effects were noted and the data that may have been influenced by these variables was not included in analyses.

Figure 3 presents the performance data for Hen 12.1. Initially, at baseline (FR 10), sensitivity was higher than specificity. As the FR requirement increased, these two measures began to converge, with sensitivity decreasing as specificity increased. The combined measure remained generally in the same range

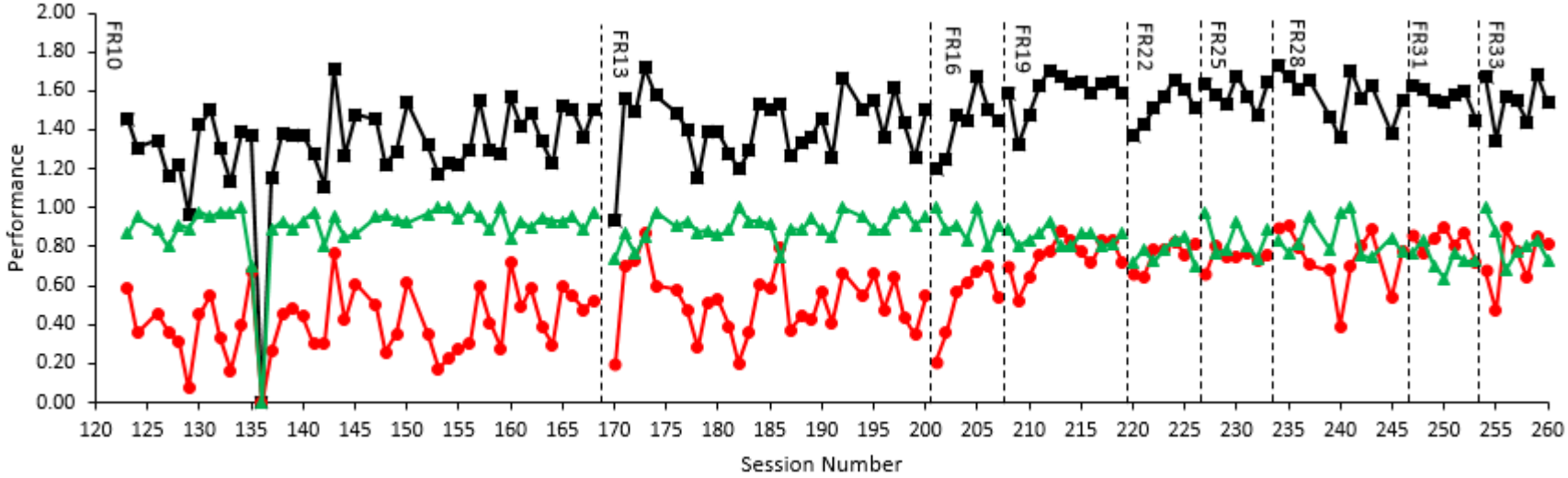

Figure 3. Performance of hen 12.1 from original baseline (FR 10) to FR 33. Specificity (red), sensitivity (green), and the combined data (black) are presented.

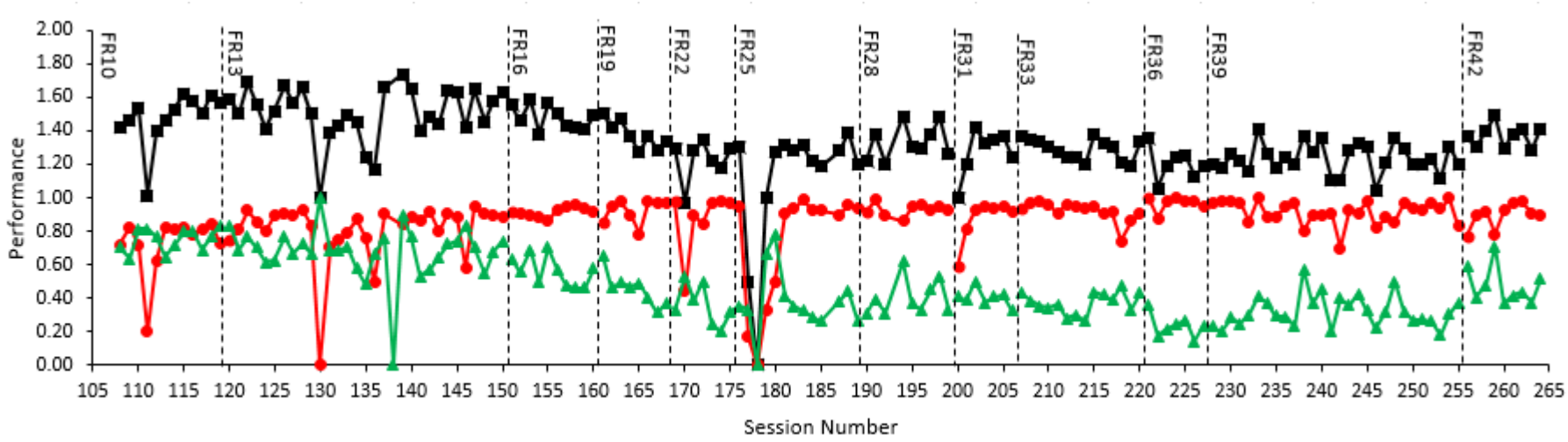

Figure 4. Performance of hen 12.2 from original baseline (FR 10) to FR 42. Specificity (red), sensitivity (green), and the combined data (black) are presented.

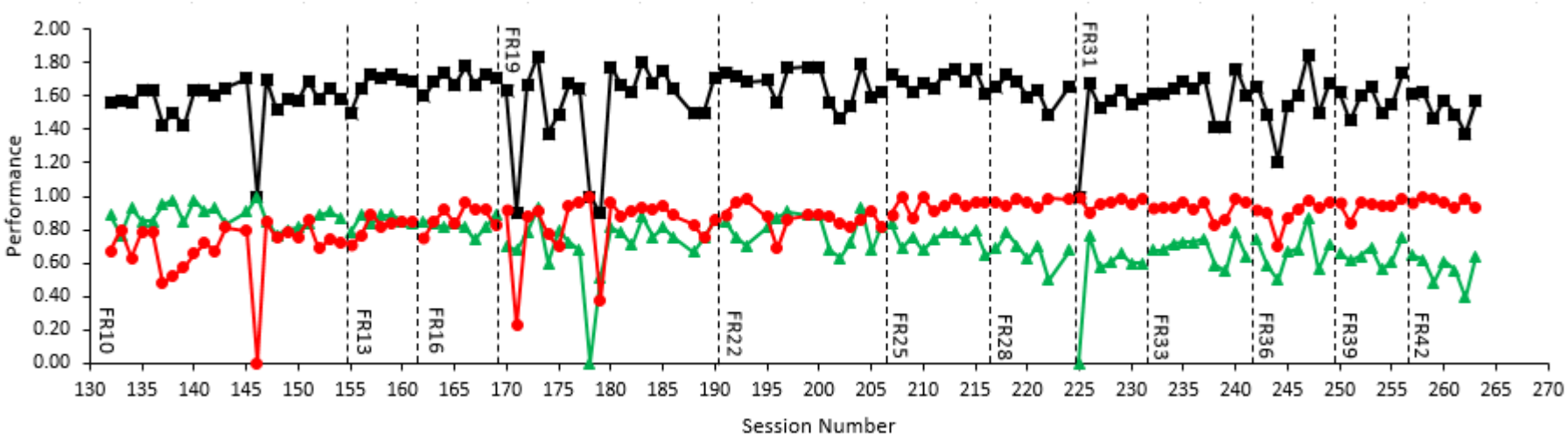

Figure 5. Performance of hen 12.3 from original baseline (FR 10) to FR 42. Specificity (red), sensitivity (green), and the combined data (black) are presented.

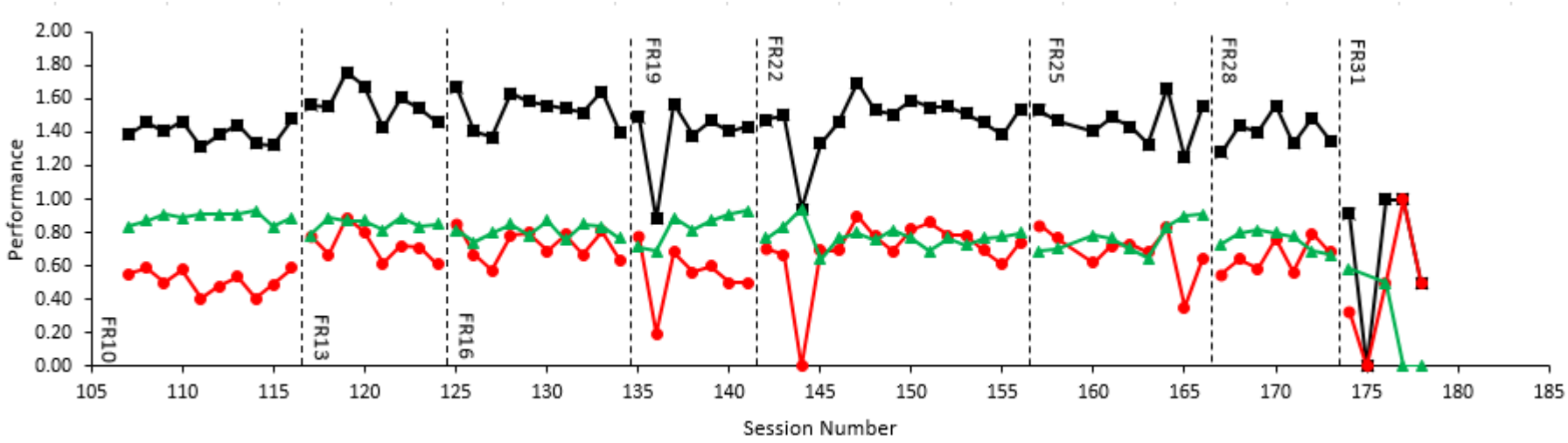

Figure 6. Performance of hen 12.4 from original baseline (FR 10) to FR 31. Specificity (red), sensitivity (green), and the combined data (black) are presented.

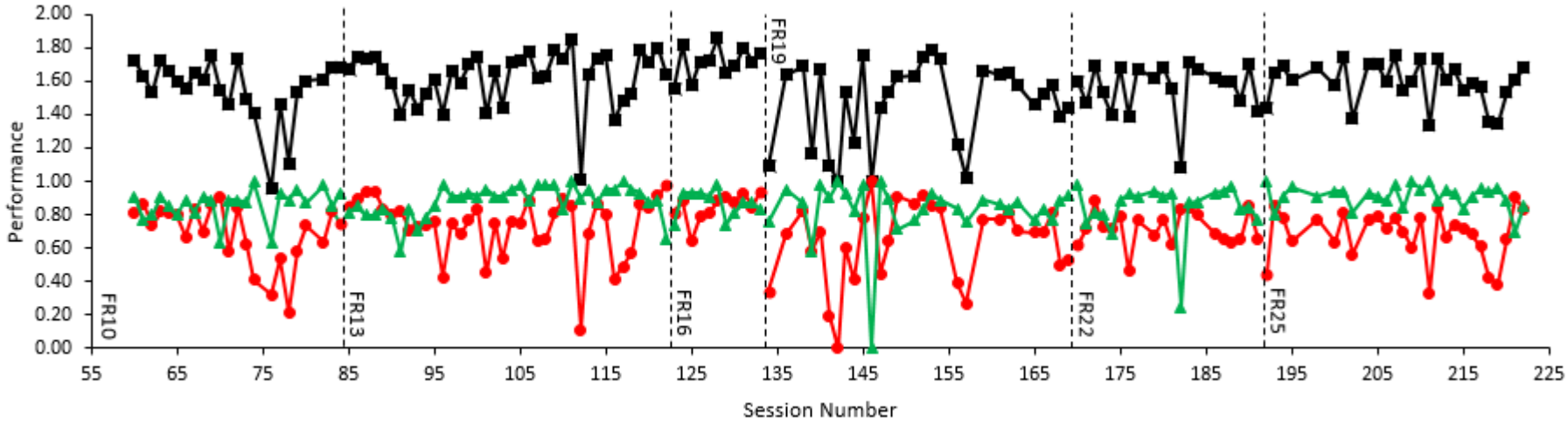

Figure 7. Performance of hen 12.5 from original baseline (FR 10) to FR 25. Specificity (red), sensitivity (green), and the combined data (black) are presented.

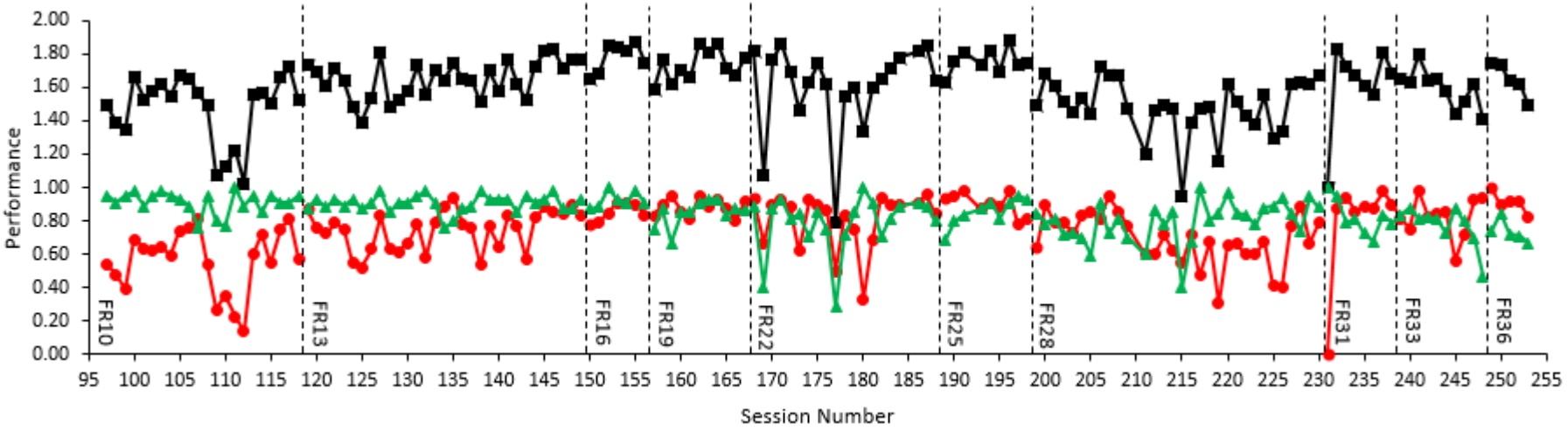

Figure 8. Performance of hen 12.6 from original baseline (FR 10) to FR 36. Specificity (red), sensitivity (green), and the combined data (black) are presented.

throughout all conditions but increased slightly as the FR value increased. Specificity was recorded higher (0.93) than sensitivity (0.80) for the first time under the FR 19 requirement, and then these two measures tended to vary within a similar range. Termination criteria were not met before baseline reinstatement.

The performance data for Hen 12.2 is displayed in Figure 4. Throughout all conditions, specificity was observed to be generally higher than sensitivity, with occasional overlap. These measures initially varied within a similar range under the original baseline (FR 10); as the FR requirement increased, sensitivity performance decreased while specificity performance increased slightly and remained high (i.e., generally close to ceiling levels). However, under the FR 42 condition (i.e., the last condition prior to baseline reinstatement), sensitivity increased slightly as specificity decreased slightly. The combined performance measure was similar to the sensitivity data, showing a decrease as FR requirement increased, with the exception of the FR 42 condition, where this measure slightly increased. Termination criteria were not met before baseline was reinstated.

Figure 5 illustrates the performance of Hen 12.3, showing that, while sensitivity was initially higher than specificity at baseline, as the FR requirement increased, sensitivity performance gradually decreased as specificity increased towards the ceiling of 1.00. This change began within early conditions (i.e., FR 13 and FR 16) where specificity and sensitivity have similar values, but the discrepancy between the two measures widened systematically, with the greatest difference in performance observed in the final condition before baseline reinstatement (FR 42). Termination criteria were not met in this condition. Performance as recorded by the combined measure was generally within the same range across conditions, showing a slight decreasing trend at the higher FR requirements (i.e., above FR 36).

The performance of Hen 12.4 is recorded in Figure 6. Initially, sensitivity was higher than specificity; however, as the FR requirement increased, specificity performance began to increase while sensitivity performance decreased slightly, until both measures were in similar ranges. Sensitivity tended to remain higher than specificity across conditions, which was particularly exemplified during the FR 19 condition, where similar discrepancies were shown between these measures as in the baseline condition; however, in the FR 22 condition, sensitivity and specificity performance once again occupied similar ranges. In the FR 25 condition, although the data for sensitivity and specificity was initially similar, as sensitivity increased, specificity decreased. This trend was reversed for the FR 28 condition, where sensitivity tended to decrease as specificity increased. Following this, in the FR 31 condition, sensitivity continued to decrease as specificity varied until termination criteria were met. The combined measure illustrated that overall performance initially increased slightly from baseline, remaining roughly within the same range for the majority of the conditions before decreasing sharply in the terminal condition (FR 31).

Figure 7 displays the performance of Hen 12.5. Generally, sensitivity was higher than specificity across all conditions, although there were some infrequent sessions when specificity was higher. As the FR requirement increased, both sensitivity and specificity varied, with sensitivity performance more stable and often close to the ceiling (i.e., performance between 0.90 and 1.00) in all conditions. Specificity was less stable, varying between decreasing and increasing trends. The combined performance measure illustrates this instability, with data varying across conditions; initially, the data appeared to trend slightly upward across baseline (FR 10) to the FR 16 condition, but combined performance was relatively low from the FR 19 condition onward. Taken individually, much of the

data from both sensitivity and specificity were within similar ranges throughout the conditions; however, when specificity was higher, sensitivity decreased, and vice versa, although this change was not always commensurate. Termination criteria were not met before baseline was reinstated.

The performance of hen 12.6 is shown in Figure 8. Initially, sensitivity was much higher than specificity at baseline, but as the FR requirement increased, these measures converged, occupying very similar ranges for the FR 16 and FR 19 conditions. Performance varied for both measures within similar ranges across the FR 19, FR 22, FR 25 and FR 28, but specificity was increasingly higher than sensitivity. In later conditions (i.e., FR 31, FR 33, and FR 36), sensitivity decreased as the FR requirement increased, and specificity rose higher than sensitivity for the majority of sessions.

Figures 9, 10, 11, 12, 13, and 14 display the performance across individual hens for the conditions from baseline reinstatement. Following baseline reinstatement, the median probe condition was repeated, and then three-quarter probe (i.e., these were individualised for each hen), before repeating the FR 13 condition. Hen 12.4 was exposed to additional, subsequent conditions as time allowed; these are described below.

Figure 9 displays the post-baseline reinstatement performance of Hen 12.1 across the four conditions of baseline, median probe, three-quarter probe, and FR 13. There was a greater discrepancy between sensitivity and specificity at lower FR values (i.e., FR 10 and FR 13) than the higher FR conditions, and sensitivity was higher than specificity for the FR 10 and FR 13 conditions. For the median probe condition (FR 22), specificity had improved, increasing to similar values as sensitivity, which had decreased slightly in comparison to baseline. For the three-quarter probe condition (FR 28), specificity increased and was equal to

or higher than sensitivity for all sessions; sensitivity was again decreased slightly relative to baseline. When the FR 13 requirement was reintroduced, specificity decreased again, while sensitivity increased.

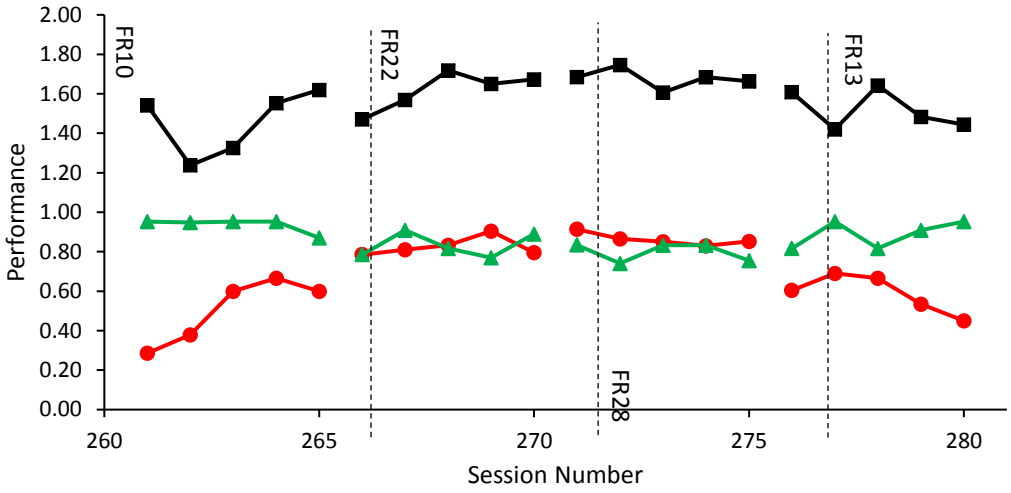

Figure 9. Performance of hen 12.1 following baseline reinstatement. Specificity (red), sensitivity (green), and the combined data (black) are presented.

Figures 10 and 11 illustrate the performance of Hens 12.2 and 12.3, respectively, following baseline reinstatement and the subsequent repeated conditions. These hens showed a similar pattern of behaviour across the reinstated conditions, and, as their maximal condition was the same (FR 42), they experienced the same repeated FR requirements per condition. Under a FR 10 schedule, sensitivity tended to be higher than specificity, but specificity trended upwards as sessions progressed. For the FR 25 and FR 33 conditions, performance was quite similar across both hens, with specificity being markedly higher than sensitivity and the range between these measures increasing. Back under a FR 13 schedule, the sensitivity and specificity measures converged again, although specificity was generally higher than sensitivity.

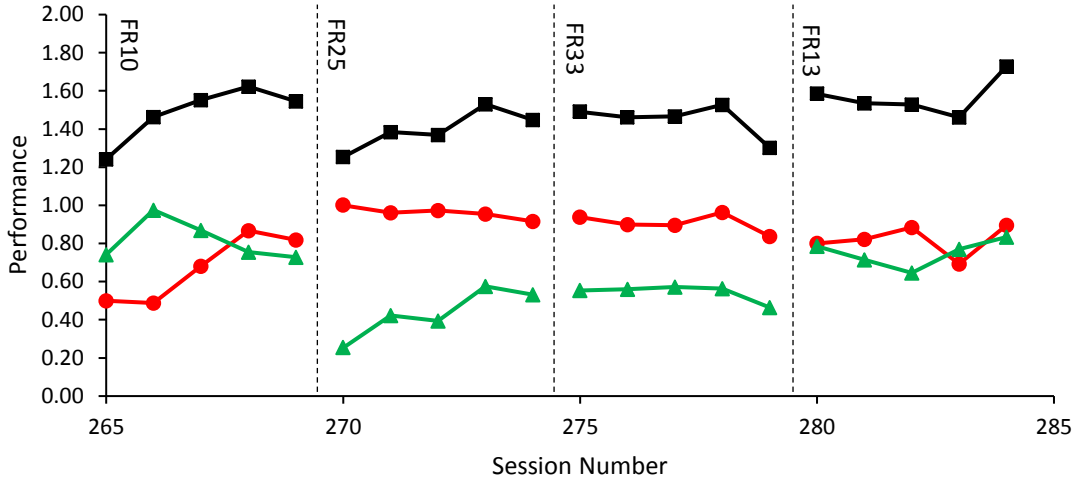

Figure 10. Performance of hen 12.2 following baseline reinstatement. Specificity (red), sensitivity (green), and the combined data (black) are presented.

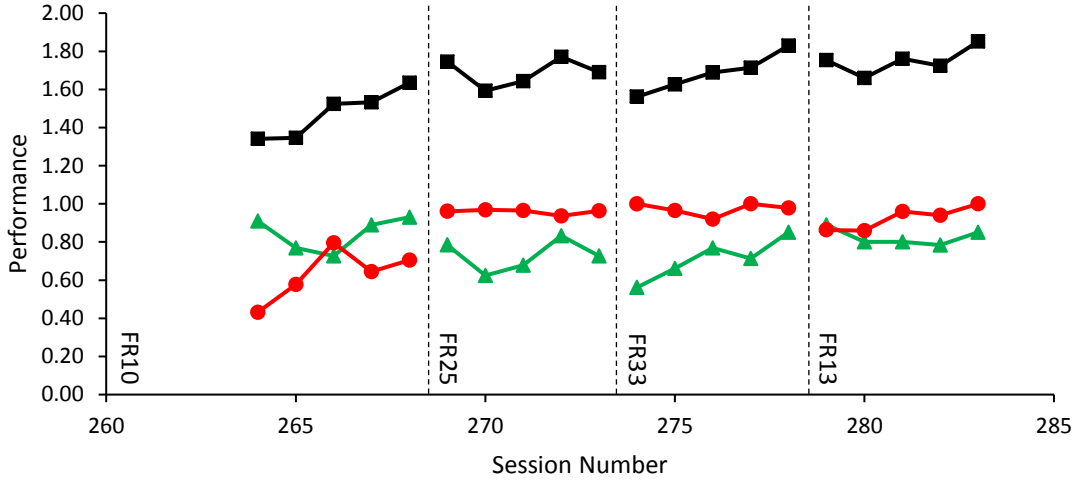

Figure 11. Performance of hen 12.3 following baseline reinstatement. Specificity (red), sensitivity (green), and the combined data (black) are presented.

Figure 12 depicts the performance of Hen 12.4 under all schedules following baseline reinstatement. For the first four conditions, this hen's performance is similar to that of 12.1; sensitivity was much higher than specificity at the lower FR values (i.e., FR 10 and FR 13), with these measures converging under the median and three-quarter probes. The discrepancy between sensitivity and specificity at lower FR values was further observed at the lowest FR requirements utilised in this study (i.e., FR 5 and FR 8). Behaviour broke down in the reinstated FR 28 condition.

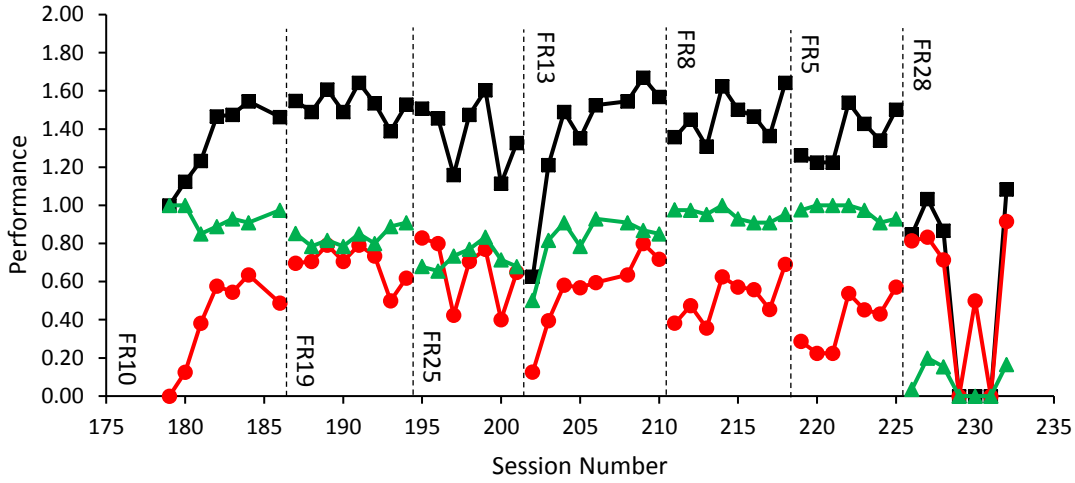

Figure 12. Performance of hen 12.4 following baseline reinstatement. Specificity (red), sensitivity (green), and the combined data (black) are presented.

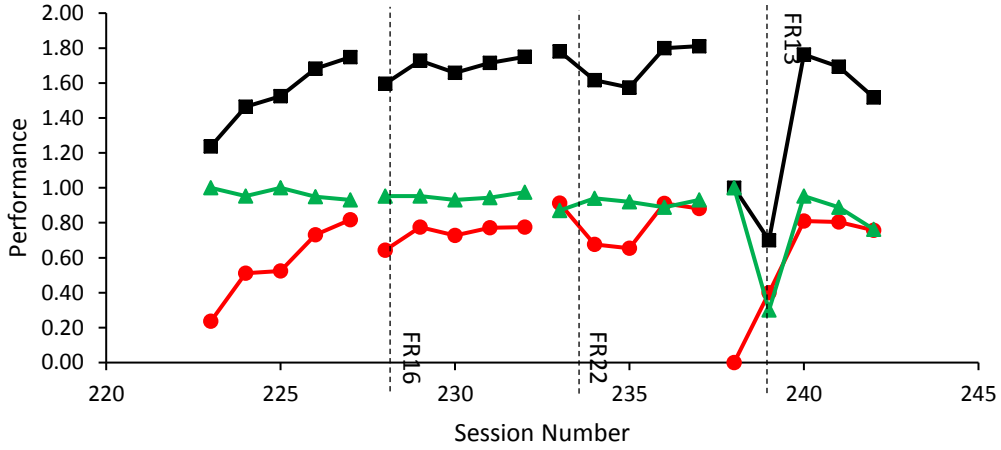

Figure 13. Performance of hen 12.5 following baseline reinstatement. Specificity (red), sensitivity (green), and the combined data (black) are presented.

Figure 13 displays the performance of Hen 12.5 across all post-baseline-reinstatement conditions. The first three conditions following baseline reinstatement demonstrate similar performance as that observed for 12.1 and 12.4; that is, with sensitivity generally observed higher than specificity under the FR 10 schedule, the two measures increasingly converge across the median and three-quarter probe conditions, and remain largely converged in the reinstated FR 13 condition as well. For all reinstated conditions, the performance of 12.5 tended to

show higher sensitivity than specificity, even when these measures were converged.

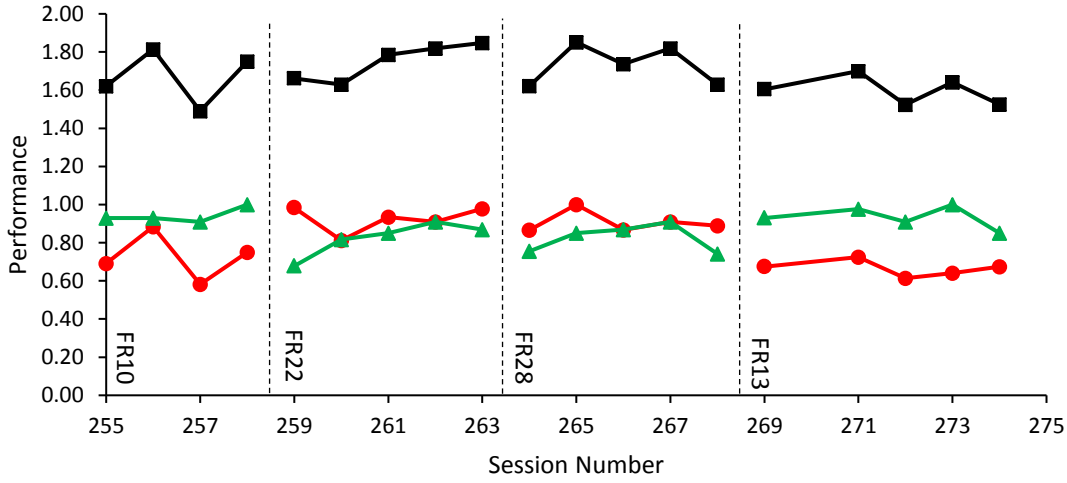

Figure 14. Performance of hen 12.6 following baseline reinstatement. Specificity (red), sensitivity (green), and the combined data (black) are presented.

Figure 14 illustrates the post-baseline-reinstatement performance of hen 12.6. Again, at lower FR values, a greater discrepancy was observed between the two measures, with sensitivity being higher than specificity. At the higher FR values of the median and three-quarter probes, sensitivity dropped slightly with specificity converging with that measure and generally rising above it, similar to the performance of 12.2 and 12.3, although 12.6's sensitivity performance was slightly different, as both measures tended to vary within relatively the same range and remain stable.

Summary Data

Summary data for each condition were generated from the raw data by calculating the mean of the last four points of each condition where the session, trial, and stability criteria for increasing the FR requirement had been met (see 'Experimental sessions'). These data were compiled for each hen in each condition. Due to some hens experiencing more FR value changes (i.e., conditions) than others, data could only be compared across all hens from the

initial baseline (FR 10) up to FR 22. The summary data were further separated into sensitivity and specificity performance across conditions; these data are represented in figures 15 and 16, respectively.

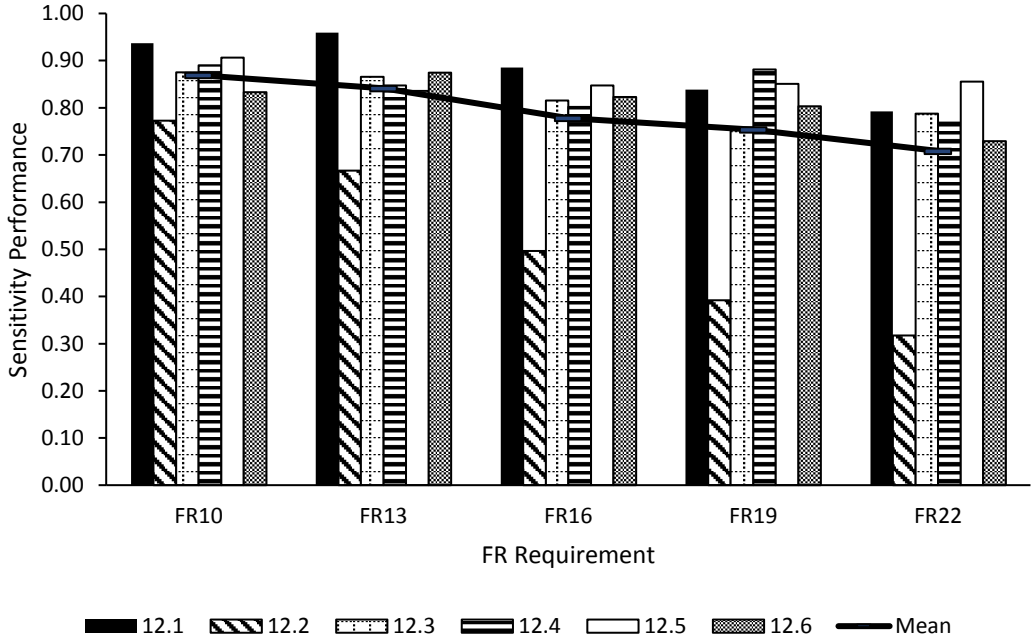

Figure 15. Sensitivity for each hen across all common FR requirements and mean sensitivity across hens in each condition.

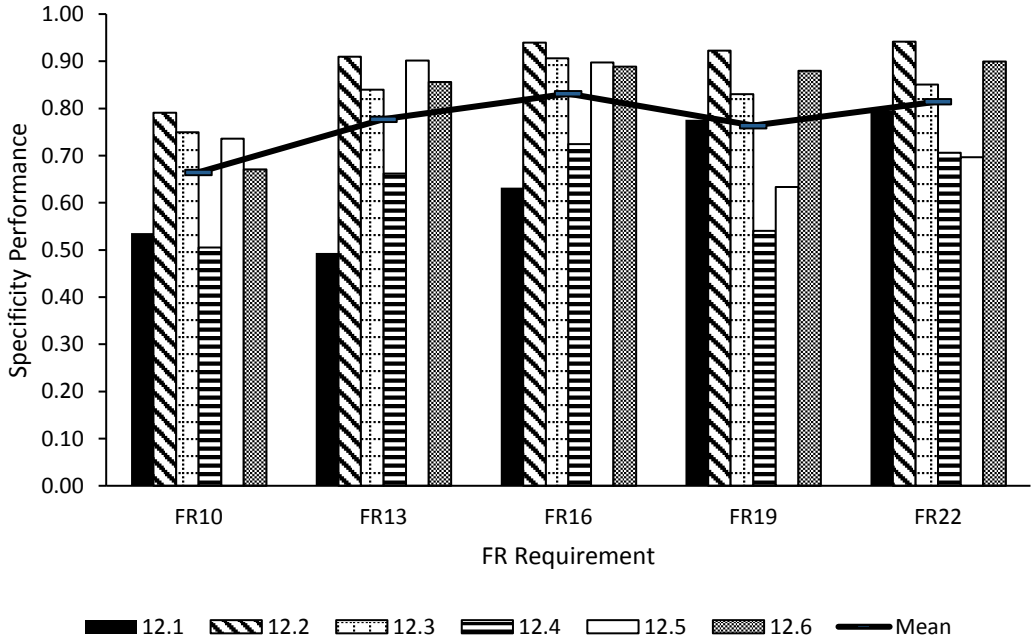

Figure 16. Specificity for each hen across all common FR requirements and mean specificity across hens in each condition.

It can be seen in Figure 15 that sensitivity performance was initially quite high for all hens, with a mean of 0.87 across subjects. Although the pattern and degree varies individually across hens, overall, sensitivity performance either remained stable or tended to slightly decrease as FR requirement increased (i.e., means of 0.84, 0.78, and 0.75 for the FR 13, FR 16 and FR 19 conditions, respectively), falling to a mean across hens of 0.71 at FR 22. However, when examining individual performance, there were idiosyncratic differences between the hens, with 12.1 initially increasing in sensitivity under the FR 13 condition before gradually decreasing for all subsequent conditions. 12.2 showed a sharp decrease across all conditions to eventually fall to a mean of 0.32 in the FR 22 condition. 12.3 and 12.4 tended to vary upwards and downwards within a similar range, decreasing slightly at the higher FR requirements. 12.5's performance was largely stable across all conditions, with the FR 16, FR 19 and FR 22 conditions all having a mean sensitivity performance of 0.85-0.86. 12.6 demonstrated a similar pattern to 12.2, with increased performance observed between the FR 10 (M = 0.83) and FR 13 (M = 0.87) conditions, with performance gradually decreasing over the remaining conditions to a mean of 0.73 under the FR 22 schedule.

Figure 16 displays the summary specificity for all hens across the common FR conditions. The pattern of specificity across these conditions presented differently for each hen, and there was no linear trend observed overall in the data. The group mean data across all hens showed increased performance across the first three conditions, with means of 0.66, 0.78, and 0.83 reported for the FR 10, FR 13, and FR 16 conditions, respectively; following the peak at FR 16, the mean specificity fell to 0.76 at FR 19, before increasing again to 0.81 under a FR 22 schedule. The individual specificity data for hens 12.2, 12.3, 12.4,

and 12.5 followed the same pattern as demonstrated in the mean data, although the specific data ranges and degree of change between conditions varied between hens. For 12.6, performance followed a similar pattern to the previous data, with the exception that the performance under the FR 22 condition was slightly higher than that of the FR 16 condition (M = 0.89, M = 0.90, respectively). For 12.1, performance decreased between the FR 10 and FR 13 conditions (M = 0.54, M = 0.49, respectively), before increasing across all remaining conditions as FR requirement increased, to a peak mean value of 0.80 under the FR 22 condition.

ROC curves for each hen, as well as the mean data, were generated to provide a visual summary of sensitivity and specificity performance which is standard in SDT research; mean data are presented in Figure 17, with the data for individual hens presented in Figure 18. For each hen, summary sensitivity data for each condition was plotted against values of (1-specificity), or false alarm rate, yielding *d'* values for each condition across hens, as presented in Table 3.

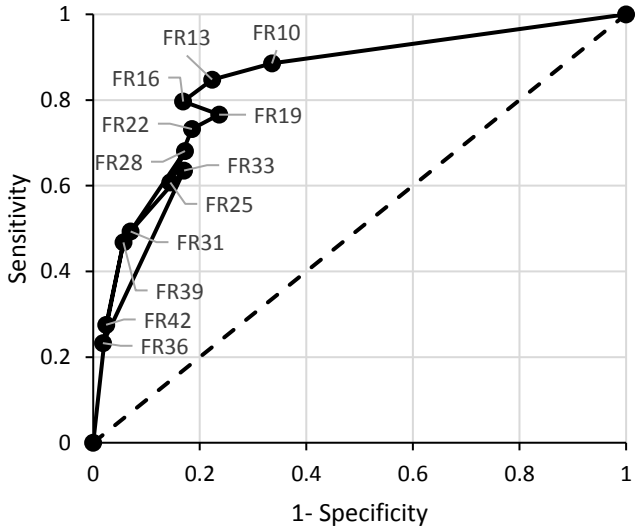

Figure 17. ROC curve generated for mean data across hens. Data points represent mean summary data per condition; data labels show the relevant condition. Dotted line represents the chance line, where performance would be related to binary guessing behaviour.

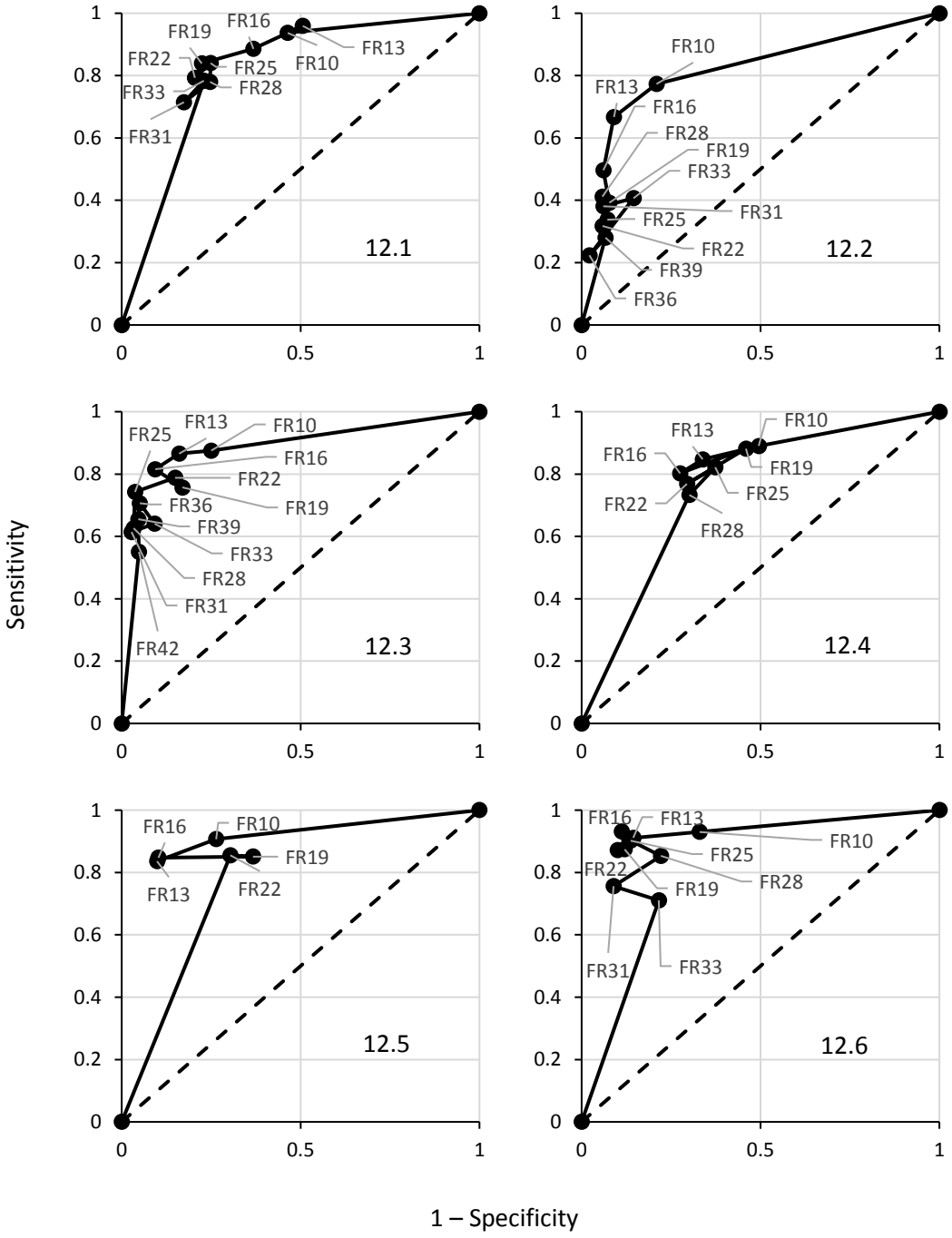

Figure 18. ROC curves generated for individual hens. Data points represent summary data per condition; data labels show the relevant condition. Dotted line represents the chance line, where performance would be related to binary guessing behaviour.

Table 3

Value of d' *Per Condition for All Hens*

	12.1	12.2	12.3	12.4	12.5	12.6
FR 10	0.472763	0.563717	0.624855	0.395366	0.642509	0.600413
FR 13	0.453073	0.576758	0.705039	0.509069	0.737341	0.766983
FR 16	0.517332	0.435917	0.721774	0.52665	0.744554	0.819325
FR 19	0.615077	0.314764	0.587411	0.421929	0.483993	0.753949
FR 22	0.58802	0.258975	0.638708	0.474276	0.551658	0.770946
FR 25	0.59163	0.265957	0.705164	0.448391		0.764936
FR 28	0.532777	0.353722	0.592818	0.432067		0.631876
FR 31	0.540552	0.320342	0.586623			0.666889
FR 33	0.553818	0.261703	0.548812			0.495043
FR 36		0.202035	0.655661			
FR 39		0.213132	0.608832			
FR 42			0.502348			

Table 4

Shapiro-Wilk Test of Normality for Sensitivity Summary Data

Condition	*p*-value
FR 10	0.12
FR 13	0.34
FR 16	0.04*
FR 19	0.004*
FR 22	0.003*

Note. Findings were considered significant at the $p < 0.05$ level; significant findings are marked with an asterisk.

In addition to these graphical analyses, statistical analyses were also run on these data to determine if there were statistically significant differences between conditions due to FR requirement changes. As Shapiro-Wilk tests of normality (as presented in Table 4) revealed some of the sensitivity data to be non-parametric, a Freidman's test was conducted, which revealed statistically significant differences in sensitivity at different FR values, $\chi^2(4) = 12.27$, $p = .015$. Further post-hoc analyses of pairwise comparisons with a Bonferroni correction are presented in Table 5; these illustrated that the only significant difference found in sensitivity was between the FR 10 (*Mdn* = 0.90) and FR 22 (*Mdn* = 0.79; p = 0.01) conditions.

Table 5

Post-hoc Pairwise Comparisons Between FR Conditions Across Hens for Sensitivity

Sample pairs	*p* (without correction)	Adjusted *p*
FR 22 - FR 19	0.47	1.00
FR 22 - FR 16	0.14	1.00
FR 22 - FR 13	0.07	0.68
FR 22 - FR 10	0.001*	0.01*
FR 19 - FR 16	0.47	1.00
FR 19 - FR 13	0.27	1.00
FR 19 - FR 10	0.01*	0.10
FR 16 - FR 13	0.72	1.00
FR 16 - FR 10	0.07	0.68
FR 13 - FR 10	0.14	1.00

Note. The adjusted *p*-values represent the findings corrected for multiple comparisons using a Bonferroni correction. Findings were considered significant at the $p < 0.05$ level; significant findings are marked with an asterisk.

Table 6

Shapiro-Wilk Test of Normality for Specificity Summary Data

Condition	*p*-value
FR 10	0.30
FR 13	0.09
FR 16	0.07
FR 19	0.51
FR 22	0.55

Note. Findings were considered significant at the $p < 0.05$ level; significant findings are marked with an asterisk.

For the specificity summary data, a Shapiro-Wilk test indicated that the assumption of normality was not violated, as presented in Table 6, and the data were further assessed for outliers by boxplot inspection (i.e., values any further than 1.5 box-lengths away from the edge of the boxes); no outliers were discovered. However, Mauchly's test of sphericity determined that the assumption of sphericity was violated, $\chi^2(9) = 20.39$, $p = .043$. Thus, a Friedman's test was run on these data, which returned a significant result, and the null hypothesis (i.e., that the specificity performance scores were distributed the same across FR conditions) was rejected, $\chi^2(4) = 13.73$, $p = .008$. Post-hoc pairwise comparisons were run using a Bonferroni correction for multiple comparisons. Table 7 presents the results of these post-hoc analyses, illustrating that two significant differences were found between the baseline (FR 10) condition and the FR 22 condition ($p = 0.047$), as well as between baseline and FR 16 ($p = 0.019$).

Table 7

Post-hoc Pairwise Comparisons Between FR Conditions Across Hens for Specificity

Sample pairs	*p* (without correction)	Adjusted *p*
FR 22 - FR 19	0.83	0.83
FR 22 - FR 16	0.78	1.00
FR 22 - FR 13	0.10	1.00
FR 22 - FR 10	0.005*	0.047*
FR 19 - FR 16	0.05	0.45
FR 19 - FR 13	0.93	1.00
FR 19 - FR 10	0.27	1.00
FR 16 - FR 13	0.06	0.55
FR 16 - FR 10	0.002*	0.019*
FR 13 - FR 10	0.24	1.00

Note. The adjusted *p*-values represent the findings corrected for multiple comparisons using a Bonferroni correction. Findings were considered significant at the $p < 0.05$ level; significant findings are marked with an asterisk.

Further regression analyses were conducted on the summary data for all hens. Visual inspection of the data suggested a linear relationship between sensitivity performance and FR requirement, and a Durbin-Watson statistic of 2.25 determined the independence of residuals, with no problematic outliers detected. However, visual inspection of a scatterplot of the standardized residuals against the standardised predicted value suggested that the assumption of homoscedasticity had been violated; this result was confirmed statistically with a Park test ($p = 0.024$).

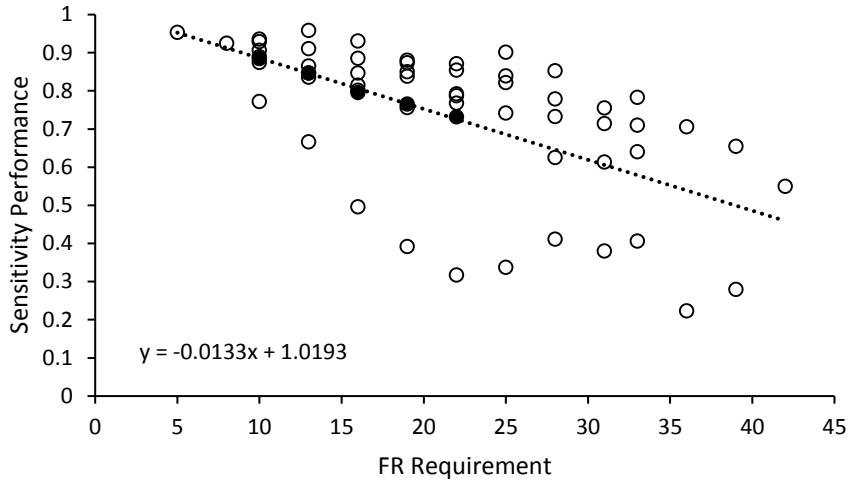

Figure 19. Linear regression of sensitivity performance across all FR requirements and all hens utilised in this study. Open circles represent summary performance of individual hens. Filled circles represent the means across all hens for the common FR conditions. The dotted line is the regression line.

A weighted least squares regression was conducted, with cases weighted such that log-likelihood function was maximised for optimal power value; subsequent graphical analysis of the weighted data confirmed homoscedasticity and approximate normality in the distribution of the residuals. FR value variation accounted for 17.9% of the variance in sensitivity performance, with a large effect size as related to Cohen's (1988) guidelines (adjusted R^2 = 15%), $F(1,28) = 14.39$, $p = 0.02$. FR requirement was found to significantly predict sensitivity performance, and a negative linear relationship was observed between the populations of these two variables, $\beta = -0.42$, $t(28) = -2.47$, $p = 0.02$, 95% CI [-0.025, -0.002]. This regression analysis yielded a predictive model for sensitivity performance, such that:

$$S = 1.019315 + (-0.013398 * FR) \qquad (1)$$

where *S* indicates predicted mean sensitivity performance and *FR* refers to a particular FR requirement. Equation 1 was used to model predicted sensitivity performance for the other FR values utilised in the experimental phases of this study; the results of these calculations are presented graphically in Figure 19. While this figure presents the summary data for all conditions, only the data from the FR 10 to FR 22 range were included in the regression analyses, as not every hen completed the other FR requirements.

A linear regression analysis was also conducted on the specificity summary data. Visual inspection of the data suggested the assumption of linearity between specificity and FR requirement was not violated, although this relationship appeared to be approximately horizontal. There were no outliers detected, and a Durbin-Watson statistic of 2.07 confirmed independence of residuals. From graphical analysis, the assumption of homoscedasticity was not violated; this was confirmed statistically using a Park test ($p = 0.97$). Graphical analysis of a normal probability plot suggested that the residuals did not violate the assumption of normality. Although the relationship was not statistically significant, systematic variance of the FR requirement accounted for 9.1% of the variance in specificity performance in the population, with a medium effect size (adjusted $R^2 = 5.9\%$). FR requirement was not found to significantly predict specificity performance, and a linear relationship was not found between these two variables, $\beta = 0.30$, $t(28) = 1.68$, $p = 0.10$. Subsequent exploratory analyses of curve estimation were run on this data to see if a polynomial or nonlinear model could provide a better explanation of the variance in specificity performance as related to FR requirement; however, none of these models appeared to be a good fit for the data, as presented in Table 8.

Table 8

Curve Estimation Models of Summary Specificity Data

Model	R^2	*p*-value
Linear	0.09	0.10
Logarithmic	0.11	0.08
Inverse	0.12	0.06
Quadratic	0.14	0.14
Cubic	0.14	0.14
Compound	0.10	0.10
Power	0.11	0.08
S	0.12	0.06
Growth	0.10	0.10
Exponential	0.10	0.10
Logistic	0.09	0.11

Note. Findings were considered significant at the $p < 0.05$ level.

Reinstatement Data

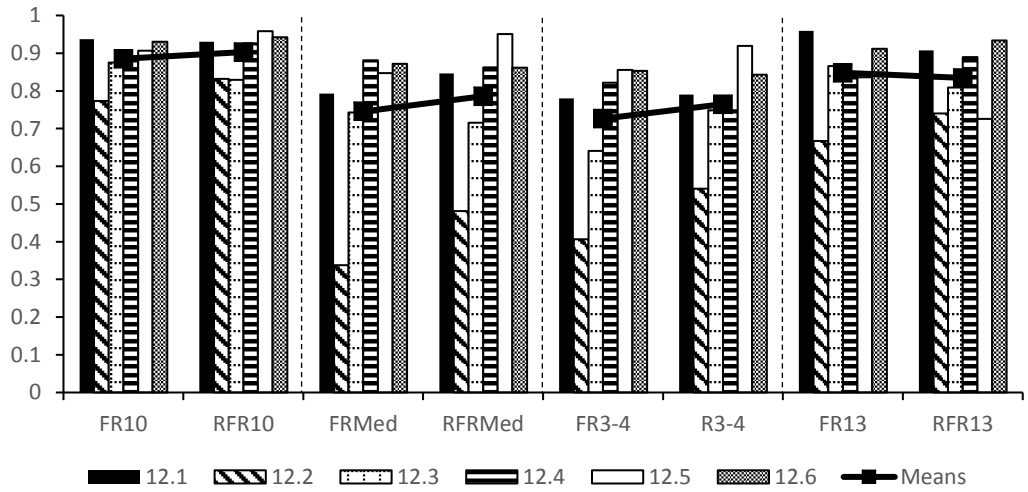

Figure 20. Comparison of sensitivity data between the original conditions and the reinstated conditions, and mean sensitivity for all hens under these conditions.

The summary performance of hens under the pre-baseline-reinstatement conditions was compared to the corresponding post-baseline-reinstatement conditions; these data are presented graphically in Figure 20, for visual reference. Performance across the initial FR 10 condition was compared with the reinstated FR 10 condition; likewise, for the median probe, where values differed for individual hens; the three-quarter probe, where individual values again differed; and the FR 13 conditions. For the summary sensitivity data from these conditions, visual inspection of boxplots did not reveal any outliers in the difference scores between the original and reinstated conditions, and Shapiro-Wilk tests of these difference scores confirmed that the assumption of normality was not violated for any condition pairs, as presented in Table 9.

Table 9

Shapiro-Wilk Test of Normality for Sensitivity Summary Difference Scores

Condition Pairs	*p*-value
FR 10 - Reinstated FR 10	0.60
FR Med - Reinstated FR Med	0.19
FR 3-4-Reinstated FR 3-4	0.84
FR 13-Reinstated FR 13	0.72

Note. Findings were considered significant at the $p < 0.05$ level.

Thus, paired-samples *T*-tests were conducted on the summary sensitivity data across all hens. For the FR 10 conditions, although the mean of the reinstated condition was higher ($M = 0.90$, $SD = 0.06$) than the original baseline ($M = 0.89$, $SD = 0.06$), a significant difference was not found between these means, $t(5) = -0.98$, $p = 0.37$. For the median probe conditions, the original conditions had a lower mean ($M = 0.75$, $SD = 0.21$) than the reinstated conditions ($M = 0.79$, $SD = 0.17$); no significant difference was found, $t(5), = -1.48$, $p = 0.20$. The mean of the

reinstated three-quarter probe conditions (M = 0.77, SD = 0.13) was higher than that of the original three-quarter probe conditions (M = 0.73, SD = 0.17), but the difference between these means was not observed to be significant, $t(5) = -1.24$, $p = 0.27$. The direction of mean difference between the original and reinstated conditions was different for the FR 13 conditions, with the original FR 13 conditions having a higher mean (M = 0.85, SD = 0.10) than the reinstated conditions (M = 0.84, SD = 0.09), but the difference between the means was not found to be statistically significant, $t(5) = 0.53$, $p = 0.62$. Thus, for all groups, the null hypothesis (i.e., the mean difference between sensitivity performance prior to baseline reinstatement and post-baseline reinstatement was zero) was not rejected.

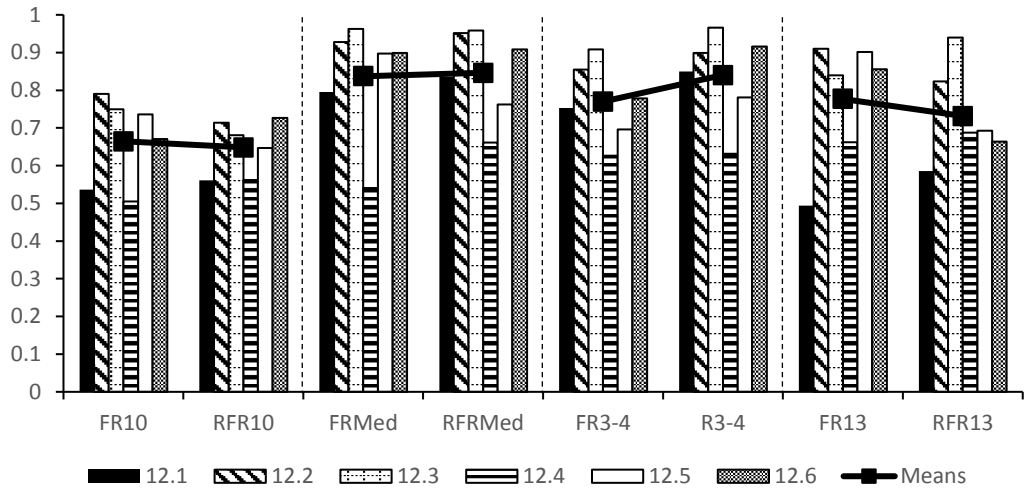

Figure 21. Comparison of sensitivity data between the original conditions and the reinstated conditions, and mean sensitivity for all hens under these conditions.

A similar process was conducted on the pre- and post-baseline-reinstatement specificity summary data across all hens; this data is presented graphically in Figure 21. Difference data were inspected visually to ensure the absence of outliers; outliers were only detected for the specificity median probe data. Shapiro-Wilk tests of normality for the difference data of the remaining three

conditions revealed that the assumption of normality was not violated, as presented in Table 10.

A paired-samples *T*-test was run on this data. For the FR 10 conditions, the mean for the original conditions ($M = 0.67$, $SD = 0.12$) was higher than that of the reinstated conditions ($M = 0.65$, $SD = 0.07$), but no significant differences were found between these means, $t(5) = 0.57$, $p = 0.59$. The mean for the original three-quarter probes ($M = 0.77$, $SD = 0.10$) was lower than the mean of the reinstated conditions ($M = 0.84$, $SD = 0.12$), and a significant difference was found between these means, $t(5) = -3.53$, $p = 0.017$, $d = 1.44$. For the FR 13 conditions, the original conditions had a higher mean ($M = 0.78$, $SD = 0.17$) than the reinstated conditions ($M = 0.73$, $SD = 0.13$), but the difference was not found to be significant, $t(5) = 0.78$, $p = 0.47$.

Table 10

Shapiro-Wilk Test of Normality for Specificity Summary Difference Scores

Condition Pairs	*p*-value
FR 10 - Reinstated FR 10	0.08
FR 3-4 - Reinstated FR 3-4	1.00
FR 13 - Reinstated FR 13	0.18

Note. Findings were considered significant at the $p < 0.05$ level.

Due to the non-normality of the difference scores for the median probe conditions, a Wilcoxon Signed-rank test was conducted on these data. The median of the original conditions ($Mdn = 0.90$) was higher than that of the reinstated conditions ($Mdn = 0.88$), but this difference was not found to be statistically significant, $z = 0.67$, $p = 0.50$.

Discussion

This study investigated the effect of systematic variations in FR requirement on the sensitivity and specificity performance of hens in a signal detection theory (SDT) task using a procedure in which reinforcement was only delivered for hits (i.e., completion of a FR criterion on the stimulus key during a positive trial), and hens could always alternately advance to the next trial by completing a FR 1 key-peck on the 'advance' key (i.e., analogous to a 'no' response in a standard SDT task, or to a 'no-go' response in a go/no-go procedure). It was hypothesised that, at lower stimulus key FR requirements, there would be a tendency towards hits (i.e., sensitivity performance would be higher than specificity), but that, as FR requirements increased, sensitivity would decrease while specificity increased.

Findings

Graphical analyses. Although there were differences in individual performance, in general, the behaviour of most hens approximated the pattern predicted, where sensitivity was generally higher than specificity at baseline (i.e., FR 10), before these two measures began to converge, after which specificity performance tended to be higher than sensitivity. The degree of discrepancy between these measures at baseline, as well as the FR condition under which the measures converged, varied across individual hens. For example, Hen 12.2's convergence point was observed in the original baseline condition, while Hen 12.1's performance on the two measures was quite disparate at baseline but converged under a FR19 schedule (see Figures 3 and 4).

The graphical analyses presented preliminary indications of support for the hypothesis, although idiosyncratic differences were observed. Idiosyncratic differences in discrimination performance with birds where the effort required to

respond was varied has been observed before, although FR requirement can be considered a relatively consistent, stable measure of effort (Elsmore, 1971). These idiosyncratic differences could potentially reflect differences in sensitivity related to the individual sensory abilities and/or the decision criteria utilised by the hens, rather than the bias related to the FR variation (Macmillan, 2002); however, measurement and control of these variables was outside the scope of the present study. To provide a more useful model of the SDT performance pattern as FR requirement changed, the summary data were calculated, comparing across hens in search of an overall effect, independent of uncontrollable, idiosyncratic differences. Figures 15 and 16 present this data graphically across the FR conditions common to all hens, allowing for both a comparison of individual differences between hens, as well as a visual representation of the effect of FR variation on mean SDT performance across hens. As FR requirement increased, these figures illustrate the general tendency for sensitivity performance to decrease and specificity performance to increase, as hypothesised. These results partially contradict those obtained by Rohles (1961) and Spetch and Treit (1986) who found that accuracy improved with greater effort requirements; the specificity results support those conclusions, but the sensitivity data do not. The effect of FR requirement variation on overall accuracy (i.e., from the combined HR+CRR measure) showed differing trends for individual hens; although this data remained largely stable for most hens, as FR requirement increased, 12.1's overall accuracy also increased, but the inverse relationship was observed for 12.2. However, the procedures utilised by those researchers differed greatly from that of the present study, as Rohles (1961) did not randomise stimulus conditions and Spetch and Treit (1986) utilised a delayed matching to sample design, which may have contributed to these differing results.

The reinstated conditions allowed for a comparison to the original conditions; in all cases, this performance approximated that of the corresponding original condition, with only some hens showing slight differences. For 12.1, changes in performance across the reinstated conditions were relatively the same as the changes observed prior to baseline reinstatement, in that there was a greater discrepancy between the values of the sensitivity and specificity measures at lower FR values, with these measures converging from FR 19 and upward; although, specificity was observed as higher than sensitivity more often during the reinstated FR 28 condition than the initial FR 28. For hens 12.2 and 12.3, again, this performance roughly matched the corresponding pre-baseline reinstatement conditions for each hen, although specificity was higher than sensitivity for 12.3 in the reinstated FR 13 condition; for this hen in the original FR 13 condition, specificity tended to vary within the same range as sensitivity, but not higher. For 12.4, performance for all of these first four, repeated conditions was similar to the performance observed under the same FR requirements in the original conditions. One final reinstated condition, FR 28, showed a similar breakdown in performance as observed in the original condition in which termination criteria was met (i.e., FR 31). Interestingly, this occurred one condition earlier in the reinstated condition, but this could have been impacted by the sudden increase in FR requirement from FR 5 to FR 28 (i.e., more of a contrast effect than an order effect). Hen 12.5's performance for the reinstated conditions was similar to that seen under the original conditions, with the exception of the reinstated FR 13 condition, which was slightly more variable on the sensitivity measure than was observed in the original FR 13 condition. Hen 12.6's performance for the reinstated conditions corresponded to the data gathered from the original conditions, where sensitivity was observed higher than specificity at FR values of

10 and 13, but increasingly converged and eventually fell below specificity levels at higher FR requirements, although sensitivity performance tended to be lower in the repeated FR 28 condition than the original.

Though there were some slight differences observed, from the initial graphical analyses, these differences did not appear to be significant. An advantage provided by these comparisons is that they allow conclusions to be drawn about the effect of the overall order of FR requirements. As performance did not appear to change significantly from the original to the reinstated conditions, it appeared that SDT performance was not significantly affected by the order in which FR requirements were implemented; that is, the changes seen in performance do not seem to be an artefact of merely the stepwise progression through FR values. FR requirement variation appeared to have an effect on SDT performance and this effect was reproducible.

Statistical analyses. Initial statistical analyses sought to compare the means of the FR conditions against each other. Friedman's tests and subsequent post-hoc analyses revealed that, across all hens, mean sensitivity performance was significantly less accurate at FR 22 than baseline, and that mean specificity performance was significantly more accurate in both FR 22 and FR 16 than baseline; however, no other significant mean differences were found. These, along with the graphical results, suggested that there might be a linear model that could better account for performance changes across FR variations, so regression analyses were conducted.

A weighted least squares regression illustrated the effect of FR requirement on sensitivity performance, with FR requirement variation significantly predicting sensitivity performance, such that increasing the FR requirement by one resulted in a decrease in mean sensitivity performance of

0.013, 95% CI [-0.025, -0.002]. That is, increasing the FR requirement above baseline resulted in systematic reductions in sensitivity (i.e., the 'yes' response occurring less when the S+ was actually present). Conversely, this meant that increasing the FR requirement resulted in the 'no-go' response occurring more when the S+ was present (i.e., the red key being pecked prior to reaching the FR requirement on the stimulus key), even though no reinforcement was delivered for this response. This fits with the conceptualisation of effort in SDT performance as a response cost (Roper & Zentall, 1999) and as a biasing variable with respect to the generalised matching law (GML; Baum, 1974; Davison & Tustin, 1978; Reed & Martens, 2008). As sensitivity decreased with increasing FR requirements, these data did not support the within-trial contrast (WTC) hypothesis, where it would be predicted that 'yes' responses leading to reinforcement would be valued more highly (i.e., there would be a bias towards 'yes' responses) when greater effort was expended before reinforcer delivery (Zentall, 2013). However, the WTC hypothesis was formulated in relation to two alternatives which both lead to reinforcement; the procedure used in this study was not equivalent, as only hits were reinforced, which could contribute to the differing findings. The findings of the present study were similar to Roper and Zentall's (1999) results, where increasing response cost caused preference to change between alternatives, with systematic reversals; varying FR requirement resulted in systematic changes in the tendency to respond 'yes' or 'no' when the S+ was present, even when the FR condition changes occurred out of the original stepwise order, as exemplified in the data from the reinstated conditions.

Although this clear linear relationship was revealed between sensitivity performance and FR requirement, a similar result was not found for specificity performance. While specificity performance increased (i.e., became more

accurate) as FR requirement increased, the linear relationship was not statistically significant, and none of the other regression models appeared to better explain the variance. One possible reason for this is that performance ceiling effects may have obscured a potential, underlying trend; that is, from baseline to FR 16, it appeared there could be a linear relationship between specificity performance and FR requirement (see Figure 16). However, from FR 16 onwards, mean specificity performance varied within relatively the same range, having high accuracy, due to the individual performance of some hens (e.g., 12.2 and 12.6 in particular) being very close to perfect (1.00). This theory was given additional support from 12.4's data at FR requirements lower than 10 (i.e., FR 5 and FR 8), which indicated a direct relationship between decreasing FR requirements and decreasing specificity performance, although only one hen was exposed to these lower FR requirements.

Statistical comparisons of performance between the original and reinstated conditions confirmed the conclusions drawn from the graphical analysis (i.e., that there were no significant differences) for all conditions but the three-quarter probe for specificity performance, where a statistically significant mean difference between the original and reinstated conditions was indicated, with performance in the reinstated conditions having a significantly higher mean (i.e., specificity was more accurate under these conditions). Given this result, as well as the lack of an appropriate predictive model for the specificity data, it suggests that there may be some extraneous variable contributing to specificity performance that was unaccounted for, and uncontrolled, in the present study. This result may be an artefact of the procedural design using concurrent schedules; that is, fluctuations in FR requirement and amount of reinforcement contacted (i.e., as per natural contingencies, only correct 'yes' responses were reinforced, and incorrect responses were not penalised, which could increase bias towards the key

providing reinforcement; Kamil et al., 1985; Voss et al., 1993), but the parameters of this relationship are not yet fully explored.

Implications

I have demonstrated that FR requirement plays a key role in determining SDT performance in a go/no-go task, and that the role of effort within a SDT paradigm is mathematically quantifiable. Therefore, future research in these areas should take response requirements into account when analysing data. There is also an implication for research design: it has been demonstrated that observing high levels of accuracy for both sensitivity and specificity is possible under an SDT paradigm where only hits are reinforced; high sensitivity is more likely seen at lower FR values, while high specificity is more likely seen at higher FR values, although these measures tend to converge at some point between these extremes. Researchers in this field, whether experimental or applied, should consider the most desirable outcome as related to their research questions, and can attempt to maximise performance on the relevant measure. In some cases, researchers will conceivably be able to select a FR requirement within the 'convergence band' of these two measures (i.e., where sensitivity and specificity converge) which provides a relatively good, accurate measure of both.

An example of a case where sensitivity performance would be most important, and thus, lower FR/effort requirements may be desired, is the work of Poling and colleagues in landmine detection, where the benefit of correctly identifying all landmines within an area (i.e., harm reduction, saving lives and allowing people to resettle on their land) could be said to outweigh the cost of an increased occurrence of false alarms (i.e., manpower hours in human metal detection; Poling et al., 2010; Poling et al., 2011a). An example of a case where having a higher specificity may be more desirable is in the studies of animal prey

detection. A hungry animal can afford to miss some prey (i.e., moving on instead of attacking requires no additional energy expenditure), but false alarms can waste precious calories by having an 'attack' response be performed when there is no prey to be caught; although, too many misses in a row could be equally problematic in this situation. Thus, researchers should be pragmatic when selecting a FR requirement for use under an SDT paradigm, based on their research questions and study design; the present study allows for increasingly informed decisions to be made in these cases.

Limitations

There were several limitations to this study in terms of the study design, as well as the generalisability of the findings. The idiosyncratic differences observed in these hens may present differently in other species, including humans; it is not certain whether these results, especially the model for mean sensitivity performance, would be obtained with other species. The experimental experience across hens was not equalised for this study, which may account for some of the individual differences.

There were pragmatic considerations which limited data collection in this study, in turn limiting the analysis and conclusions that can be drawn. This was largely due to the fact that not all of the hens experienced every condition, so the comparisons across hens were limited to five out of all of the FR conditions because of missing data. Because of the nature of the study (i.e., FR increasing incrementally) and analyses (Friedman's tests and regression), it was not possible to interpolate the missing data. In addition, because of the closure of the laboratory, only one hen was able to meet the termination criteria; it is unclear what the performance of the other hens would have been at higher FR requirements, and it is unknown when their performance would have broken down

or changed in a manner that would have affected the conclusions drawn above. In addition, although it was of interest to observe performance at both very high and low FR requirements (i.e., extremes), pragmatic concerns meant that experimental data was only gathered at very low FR requirements for one hen (12.4), and the highest FR requirement was not as extreme as had been observed in other research (e.g., FR 64; Elsmore, 1971) and was only reached by two hens (12.2 and 12.3). Also, to better ensure that performance is not affected by overall order effects of FR requirement presentation, it may have been more useful for all of the conditions to be reinstated in a randomised order across hens, had time allowed.

Some statistical limitations also apply to this study. Although a linear model was generated for mean sensitivity performance, these data were technically not suitable for this type of analysis, because they were bounded. Because of the way that data were aggregated as proportional data ranging from the possible values of 0.00 to 1.00, a linear model of this data will sometimes return results that don't make practical sense (i.e., values outside the bounds of 0.00-1.00). However, there is some evidence that linear regression for proportion data, when interpreted with care, are an appropriate statistical measure (Ferrari & Cornelli, 2016), although logistic regression is usually employed for proportion data, but the data set up of the present study did not allow for this. Another limitation was the lack of sufficient explanation for the changes observed in the mean specificity data (i.e., no appropriate model could be generated).

Considerations for Future Research

Future research should focus on widening the applicability of the results; that is, foremost, researchers could attempt to find a model for specificity data. Theoretically, this would allow for researchers whose investigations place more value on specificity than sensitivity to select an optimal FR value with more

certainty. It would also be interesting to compare the effect of variations in reinforcement rate, magnitude or probability on SDT performance with the effect of effort, and to attempt to quantify the interaction of the two effects on sensitivity and specificity. Future research could also attempt to generalise these findings across different species; the effect of effort variation on human SDT performance as compared to that of animals would be an interesting comparison. In addition, performance relating to the systematic variation of different types of effort could be investigated, extending the work of Elsmore (1971). For example, FR requirement variation could be compared to task difficulty, as well as force (i.e., providing sub-criterion responses were accounted for; Pinkston & Libman, 2017), as measures of effort under an SDT paradigm.

Conclusion

The main finding of this study was the presence of a negative linear relationship between FR requirement variation and sensitivity performance as described above, in line with the interpretation that 'effort' in the form of FR requirement acts as a biasing variable and may be quantified as per the GML. Although a statistically significant result was not obtained for the specificity data, this data did show a general graphical trend of increasing directly with FR requirement increases. These findings have implications for a wider range of studies and applications involving go/no-go SDT procedures.

References

Abdi, H. (2007). Signal detection theory. In N. Salkind (Ed.), *Encyclopedia of measurement and statistics* (Vol. 3, pp. 887-889). Thousand Oaks, CA: Sage. https://doi.org/10.4135/9781412952644.n405

Alling, K., & Poling, A. (1995). The effects of differing response-force requirements on fixed-ratio responding of rats. *Journal of the Experimental Analysis of Behavior, 63*(3), 331-346. https://doi.org/10.1901/jeab.1995.63-331

Baum, W. M. (1974). On two types of deviation from the matching law: bias and undermatching. *Journal of the Experimental Analysis of Behavior, 22*(1), 231-242. https://doi.org/10.1901/jeab.1974.22-231

Billington, E., & DiTommaso, N. M. (2003). Demonstrations and applications of the matching law in education. *Journal of Behavioral Education, 12*(2), 91-104. https://doi.org/10.1023/A:1023881502494

Boneau, C. A., & Cole, J. L. (1967). Decision theory, the pigeon, and the psychophysical function. *Psychological Review, 74*(2), 123-135. Retrieved from Retrieved from http://psycnet.apa.org/journals/rev/

Blough, D. S. (2001). Some contributions of signal detection theory to the analysis of stimulus control in animals. *Behavioural Processes, 54*(1-3), 127-136. https://doi.org/10.1016/S0376-6357(01)00154-1

Campbell, R. A. (1965). Thresholds and (or?) signal detection theory [Letter to the editor]. *Journal of Speech, Language, and Hearing Research, 8*(1), 97-98. https://doi.org/10.1044/jshr.0801.97

Chung, S-H. (1965). Effects of effort on response rate. *Journal of the Experimental Analysis of Behavior, 8*(1), 1-7. https://doi.org/10.1901/jeab.1965.8-1

Clement, T. S., Feltus, J. R., Kaiser, D. H., & Zentall, T. R. (2000). "Work ethic" in pigeons: Reward value is directly related to the effort or time required to obtain the reward. *Psychonomic Bulletin & Review, 7*(1), 100-106. https://doi.org/10.3758/BF03210727

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences.* New York. NY: Academic.

Davison, M. C., & Tustin, R. D. (1978). The relation between the generalized matching law and signal-detection theory. *Journal of the Experimental Analysis of Behavior, 29*(2), 331-336. https://10.1901/jeab.1978.29-331

Elsmore, T. F. (1971). Effects of response effort on discrimination performance. *The Psychological Record, 21*(1), 17-24.

Ferrari, A., & Cornelli, M. (2016). A comparison of methods for the analysis of binomial clustered outcomes in behavioral research. *Journal of Neuroscience Methods, 274*(1), 131-140. https://doi.org/10.1016/j.jneumeth.2016.10.005

Friman, P. C., & Poling, A. (1995). Making life easier with effort: Basic findings and applied research on response effort. *Journal of Applied Behavior Analysis, 28*(4), 583-590. https://doi.org/10.1901/jaba.1995.28-583

Getty, T., Kamil, A. C., & Real, P. G. (1987). Signal detection theory and foraging for cryptic or mimetic prey. In A. C. Kamil, J. R. Krebs, & H. R. Pulliam (Eds.), *Foraging behavior* (pp. 525-548). New York, NY: Springer.

Gonzalez, R. C., Bainbridge, P., & Bitterman, M. E. (1966). Discrete-trials lever pressing in the rat as a function of pattern of reinforcement, effortfulness of response, and amount of reward. *Journal of Comparative and Physiological Psychology*, *61*(1), 110-122. https://doi.org/10.1037/h0022878

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.

Harries, P., Yang, H., Davies, M., Gilhooly, M., Gilhooly, K., & Thompson, C. (2014). Identifying and enhancing risk thresholds in the detection of elder financial abuse: A signal detection analysis of professionals' decision making. *BMC Medical Education, 14*, 1044-1055. https://doi.org/10.1186/s12909-014-0268-z

Herrnstein, R. J. (1958). A conjunctive schedule of reinforcement. *Journal of the Experimental Analysis of Behavior, 1*(1), 15-24. https://doi.org/10.901/jeab.1958.1-15

Herrnstein, R. J. (1961). Relative and absolute strength of responses as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior, 4*(3), 267–72. https://doi.org/10.1901/jeab.1961.4-267

Hirzel, T. (2017). PWM. Retrieved from https://www.arduino.cc/en/Tutorial/PWM

Horner, R. H., & Day, H. M. (1991). The effects of response efficiency on functionally equivalent competing behaviours. *Journal of Applied Behavior Analysis, 24*(4), 719-732. https://doi.org/10.1901/jaba.1991.24-719

Jenkins, H. M. (1965). Measurement of stimulus control during discriminative operant conditioning. *Psychological Bulletin, 64*(5), 365-376. https://doi.org/10.1037/h0022537

Kamil, A. C., Lindstrom, F., & Peters, J. (1985). The detection of cryptic prey by blue jays (*Cyanocitta cristata)* I: The effects of travel time. *Animal Behaviour, 33*(4), 1068-1079. https://doi.org/10.1016/S0003-3472(85)80165-2

Kamil, A. C., Yoerg, S. I., & Clements, K. C. (1988). Rules to leave by: Patch departure in foraging blue jays. *Animal Behaviour, 36*(3), 843-853. https://doi.org/10.1016/S0003-3472(88)80167-2

Lattal, K. A. (1979). Reinforcement contingencies as discriminative stimuli: II. Effects of changes in stimulus probability. *Journal of the Experimental Analysis of Behavior, 31*(1), 15-22. https://doi.org/10.1901/jeab.1979.31-15

Lydall, E. S., Gilmour, G., & Dwyer, D. M. (2010). Rats place greater value on rewards produced by high effort: An animal analogue of the "effort justification" effect. *Journal of Experimental Social Psychology, 46*(6), 1134-1137. https://doi.org/10.1016/j.jesp.2010.05.011

Macmillan, N. A. (2002). Signal detection theory. In H. Pashler, & J. Wixted (Eds.), *Stevens' handbook of experimental psychology: Methodology in experimental psychology* (3rd ed., Vol. 4, pp. 43-90). New York, NY: John Wiley & Sons.

Malmberg, K. J. (2002). On the form of ROCs constructed from confidence ratings. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(2), 380-387. https://doi.org/10.1037//0278-7393.28.2.380

McCarthy, D. (1981). Toward a unification of psychophysical and behavioural research. *New Zealand Psychologist, 10*(1), 2-14. Retrieved from http://www.psychology.org.nz

McCarthy, D., & Davison, M. (1979). Signal probability, reinforcement, and signal detection. *Journal of the Experimental Analysis of Behavior, 32*(3), 373-386. https://doi.org/10.1901/jeab.1979.32-373

McCarthy, D., & Davison, M. (1981). Towards a behavioral theory of bias in signal detection. *Perception & Psychophysics, 29*(4), 371-382. https://doi.org/10.3758/BF03207347

McFall, R. M., & Treat, T. A. (1999). Quantifying the information value of clinical assessments with signal detection theory. *Annual Review of Psychology, 50*(1), 215-241. https://doi.org/10.1146/annurev.psych.50.1.215

Mumpower, J. L., & McClelland, G. H. (2014). A signal detection theory analysis of racial and ethnic disproportionality in the referral and substantiation processes of the U.S. child welfare services system. *Judgement and Decision Making, 9*(2), 114-128. Retrieved from http://journal.sjdm.org/

Nevin, J. A. (1965). Decision theory studies of discrimination in animals. *Science, 150*(3699), 1057. https://doi.org/10.1126/science.150.3699.1057

Nevin, J. A. (1969). Signal detection theory and operant behavior. [Review of the book *Signal detection theory and psychophysics,* by D. M. Green & J. A. Swets]. *Journal of the Experimental Analysis of Behaviour, 12*(3), 475-480. https://doi.org/10.1901/jeab.1969.12-475

Nishiyama, R. (2014). Response effort discounts the subjective value of rewards. *Behavioural Processes, 107*(1), 175-177. https://doi.org/10.1016/j.beproc.2014.08.002

Piazza, C. C., Roane, H. S., Keeney, K. M., Boney, B. R., & Abt, K. A. (2002). Varying response effort in the treatment of pica maintained by automatic reinforcement. *Journal of Applied Behavior Analysis, 35*(3), 233-246. https://doi.org/10.1901/jaba.2002.35-233

Pietrewicz, A. T., & Kamil, A. C. (1977). Visual detection of cryptic prey by blue jays (*Cyanocitta cristata*). *Science, 195*(4278), 580-582. https://doi.org/10.1126/science.195.4278.580

Pinkston, J. W., & Libman, B. M. (2017). Aversive functions of response effort: Fact or articfact? *Journal of the Experimental Analysis of Behavior, 108*(1), 73-96. https://doi.org/10.1002/jeab.264.

Poling, A., Weetjens, B. J., Cox, C., Beyene, N. W., Bach, H., & Sully, A. (2011a). Using trained pouched rats to detect land mines: Another victory for operant conditioning. *Journal of Applied Behavior Analysis, 44*(2), 351-355. https://doi.org/10.1901/jaba.2011.44-351

Poling, A., Weetjens, B. J., Cox, C., Beyene, N. W., Durgin, A., & Mahoney, A. (2011b). Tuberculosis detection by giant African pouched rats. *The Behavior Analyst, 34*(1), 47-54. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/

Poling, A., Weetjens, B. J., Cox, C., Beyene, N. W., & Sully, A. (2010). Using giant African pouched rats (*Cricetomys gambianus*) to detect landmines. *The Psychological Record, 60*(4), 715-728. Retrieved from http://opensiuc.lib.siu.edu/tpr/

Reed, D. D., & Martens, B. K. (2008). Sensitivity and bias under conditions of equal and unequal academic task activity. *Journal of Applied Behavior Analysis, 41*(1), 39-52. https://doi.org/10.1901/jaba.2008.41-39

Reed, G. K. (2010). Reinforcement. In C. S. Clauss-Ehlers (Ed.), *Encyclopedia of cross-cultural school psychology* (pp. 796-799). New York, NY, USA: Springer.

Reither, K., Jugheli, L., Glass, T. R., Sasamalo, M., Mhimbira, F. A., Weetjens, B. J., … Mahoney, A. (2015). Evaluation of giant African pouched rats for detection of pulmonary tuberculosis in patients from a high-endemic setting. *PLoS ONE, 10*(10), 1-13. https://doi.org/10.1371/journal.pone.0135877

Repperger, D. W., Aleva, D. L., Thomas, G., Miller, J. E., & Fullenkamp, S. C. (2007). Complexity of visual icons studied via signal detection theory. *Perceptual and Motor Skills, 105*(1), 287-298. https://doi.org/10.2466/PMS.105.1.287-298

Robin, D. A., & McNeil, M. R. (1994). The use of signal detection theory to evaluate aphasia diagnostic accuracy and clinician bias [Erratum]. *Clinical Aphasiology, 22*, 165-179. Retrieved from http://aphasiology.pitt.edu/

Rohles, F. H., Jr. (1961). The development of an instrumental skill sequence in the chimpanzee. *Journal of the Experimental Analysis of Behavior, 4*(4), 323-325. https://doi.org/10.1901/jeab.1961.4-323

Roper, K. L., & Zentall, T. R. (1999). Observing behavior in pigeons: The effect of reinforcement probability and response cost using a symmetrical choice procedure. *Learning & Motivation, 30*(3), 201-220. https://doi.org/10.1006/lmot.1999.1030

Rouder, J. N., & Morey, R. D. (2009). The nature of psychological thresholds. *Psychological Review, 116*(3), 655-660. https://doi.org/10.1037/a0016413

Shahan, T. A., & Chase, P. N. (2002). Novelty, stimulus control, and operant variability. *The Behavior Analyst, 25*(2), 175-190. https://doi.org/10.1007/BF03392056

Spetch, M. L., & Treit, D. (1986). Does effort play a role in the effect of response requirements on delayed matching to sample? *Journal of the Experimental Analysis of Behavior, 45*(1), 19-31. https://doi.org/10.1901/jeab.1986.45-19

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31*(1), 137-149. https://doi.org/10.3758/BF03207704

Tanner, W. P., Jr., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review, 61*(6), 401-409. Retrieved from http://psycnet.apa.org/journals/rev/

Tsukamoto, M., Kohara, K., & Takeuchi, K. (2017). Effects of effort and difficulty on human preference for a stimulus: Investigation of the within-trial contrast. *Learning & Behavior, 45*(2), 135-146. https://doi.org/10.3758/s13420-016-0248-8

Voss, P., McCarthy, D., & Davison, M. (1993). Stimulus control and response bias in an analogue prey-detection procedure. *Journal of the Experimental Analysis of Behavior, 60*(2), 387-413. https://doi.org/10.1901/jeab.1993.60-387

Wilson, A. N., Glassford, T. S., & Koerkenmeier, S. M. (2016). Effects of response effort on resurgence. *Behavior Analysis in Practice, 9*(2), 174-178. https://doi.org/10.1007/s40617-016-0122-3

Zentall, T. R. (2013). Animals prefer reinforcement that follows greater effort: Justification of effort or within-trial contrast? *Comparative Cognition & Behavior Reviews, 8*, 60-77. https://doi.org/10.3819/ccbr.2013.80004

Zhou, L., Goff, G. A., & Iwata, B. A. (2000). Effects of increased response effort on self-injury and object manipulation as competing responses. *Journal of Applied Behavior Analysis, 33*(1), 29-40. https://doi.org/10.1901/jaba.2000.33-29